Saisie de prescriptions médicales en langage naturel sur terminaux mobiles

Ali Can Kocabiyikoglu

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Ali Can Kocabiyikoglu

Thèse dirigée par **François PORTET**
et codirigée par **Hervé BLANCHON**

préparée au sein du **Laboratoire d'Informatique de Grenoble (LIG)**
et de **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Saisie de prescriptions médicales en langage naturel sur terminaux mobiles

Thèse soutenue publiquement le **23 Mars 2022**,
devant le jury composé de :

Mme. Christine VERDIER
Professeure, Université Grenoble Alpes, LIG, Présidente
Mme. Sophie ROSSET
Directrice de recherche, LISN, CNRS, Rapportrice
Mme. Sandra BRINGAY
Professeure, Université de Montpellier 3, LIRMM, Rapportrice
Mr. Marc CUGGIA
Professeur - Praticien Hospitalier, Université de Rennes 1, Examinateur
M. Jean-Marc BABOUCHKINE
Docteur - Président Directeur Général, Calystene, Invité
Mme. Prudence GIBERT
Docteur, CHU Grenoble Alpes, Pôle Pharmacie, Invitée
M. François PORTET
Professeur, LIG, Université Grenoble Alpes, Directeur de thèse
M. Hervé Blanchon
Maître de Conférences, HDR, LIG, Université Grenoble Alpes, Co-Directeur de thèse

Remerciements

Je tiens à remercier en premier mon directeur thèse François Portet qui m'a encadré tout au long de cette thèse, même bien avant lors de mon stage de Master 2. Je remercie également Hervé Blanchon qui a co-encadré cette thèse et qui m'a accompagné pendant ma période d'ATER à l'IUT2. Tout au long de ces trois ans et demi de thèse, ils m'ont encouragé à aller au bout de celle-ci, m'ont donné des conseils précieux mais aussi la possibilité de poursuivre mes idées tout en s'adaptant aux contextes changeants de ces trois dernières années.

Je tiens à remercier aussi Jean-Marc Babouchkine et mes collègues à Calystene qui ont partagé leurs expertises dans le domaine du biomédical. Dans la réalisation de cette thèse, je salue les efforts immenses des cliniciens du département de gérontologie du CHU de Grenoble, notamment Dr. Prudence Gibert et Professeur Gaëtan Gavazzi qui nous ont accueillis, écoutés et qui ont participé à la réalisation et à la collecte de données. Je remercie également tous les professionnels de santé qui ont participé, même parfois pendant leur pause déjeuner, à la réalisation de cette thèse malgré le contexte sanitaire très tendu à l'époque.

Je remercie bien sûr tous mes collègues au LIG. Cela fait 6 ans que je suis au Laboratoire Informatique de Grenoble, et j'y ai connu beaucoup de personnes remarquables. Je remercie en premier tous les membres de l'équipe GETALP qui m'ont accueilli chaleureusement dans cette équipe formidable. Je tiens à remercier Laurent Besacier de m'avoir accueilli au LIG pendant mon stage de fin de Master et qui m'a encouragé à continuer dans la recherche.

Je voudrais remercier également tous mes collègues et mes enseignants du Master IdL. En particulier, Olivier Kraif qui m'a appris notamment les bases de la programmation et qui m'a encadré en stage lors de ma première année de Master. Je remercie particulièrement mes collègues Doriane Simonnet, Louise Tarrade, William N. Havard, Pauline Soutrenon et Maria Taranina qui ont toujours été là pendant le Master et tout au long de cette thèse. Je suis très heureux de voir que nous avons tous terminé nos thèses ou que nous travaillons toujours dans le TAL.

Je remercie également, mes amis Can Hazır, Asiye Kurt et Emre Hergünaç et Marion Beauc qui ont été là tout au long de cette thèse, malgré les confinements et les évolutions constantes, nous sommes toujours restés en contact et cela nous a permis de garder le moral (même parfois en visio).

Je tiens à remercier également mes enseignants de la licence SDL à l'université de Rouen, en particulier Mr. Mehmet Ali Akıncı et Laurent Gosselin qui m'ont encouragé à poursuivre un master à Grenoble et qui m'ont rassuré sur mes capacités à faire une formation de TAL.

Pour terminer, je voudrais remercier mes parents, ma pacstenaire Cécile Crépin et sa famille qui ont toujours été là pour moi. Sans leur soutien moral inconditionnel, la réalisation de cette thèse n'aurait pas été possible.

Résumé

L'utilisation des technologies de l'information dans les établissements de santé permet une meilleure organisation, renforce les procédures, permet un flux d'informations continu et sécurise le processus de soins. L'un des composants faisant partie des *systèmes informatiques de santé* (SIS) est un logiciel d'aide à la prescription (LAP) qui permet de limiter les événements indésirables médicamenteux (EIM). Lorsque les médecins utilisent un logiciel pour les soins de santé, la plupart des données sont saisies manuellement dans le LAP, ce qui réduit le temps consacré aux soins. Afin de surmonter cet obstacle, nous proposons de fournir une interface en langage naturel au LAP afin que les praticiens puissent enregistrer leurs prescriptions orales sur le lieu de soins en utilisant un terminal mobile. Un tel système permettrait aux praticiens d'utiliser un LAP en déplacement et leur permettrait de se rapprocher le plus possible de la manière la plus naturelle de prescrire des médicaments.

L'approche générale que nous avons adoptée consiste à proposer un système de dialogue orienté tâche couplé à un LAP destiné aux prescripteurs. L'un des principaux défis à relever a été de concevoir un système complet de dialogue spécialisé avec des traitements majoritairement basés sur des méthodes d'apprentissage automatique profond sans donnée d'entrainement disponible. Pour contourner ce problème, nous avons proposé une méthode itérative couplant de la collecte de corpus, de la génération automatique de texte et de la modélisation par expertise. Nous présentons une modélisation conversationnelle validée auprès d'experts médicaux et une étude détaillée des caractéristiques des prescriptions médicamenteuses d'un point de vue de TALN. L'un des principaux composants d'un système de dialogue est axé sur le processus de compréhension du langage naturel (NLU), qui est abordé par une approche de *slot-filling*. Notre méthode de collecte et de génération de données a permis de créer un corpus équilibré et couvrant l'ensemble de la sémantique de prescription que nous avons définie. Ce corpus a permis l'apprentissage initial de modèle de NLU ainsi que l'amorçage du système de dialogue.

Pour valider notre approche et recueillir des données de prescriptions réalistes à l'oral, nous avons proposé une expérimentation avec le double objectif d'évaluer le système et de collecter des données. L'évaluation qui a inclus 55 personnes, dont 34 experts médicaux, a montré que les performances de NLU étaient satisfaisantes avec une F-mesure de 90% pour un taux de réussite de la tâche de plus de 75% pour les experts médicaux. Ces résultats sont comparables à ceux obtenus sur le corpus initial ce qui confirme que la démarche adoptée durant la thèse était valide. Afin de favoriser la recherche reproductible, le corpus oral aligné (parole-transcription-sémantique) comprenant plus de 4 heures d'enregistrement sera diffusé à la communauté.

Cette thèse a été effectuée dans le cadre d'une collaboration CIFRE (convention industrielle de formation par la recherche) entre la société Calystene SA et Laboratoire Informatique de Grenoble (LIG).

Mots clés : Systèmes de dialogue oral, compréhension du langage naturel (NLU), informatique biomédicale, TAL biomédical

Abstract

The use of information technology in health care facilities allows for better organization, strengthens procedures, allows for a continuous information tranmission and makes the care process safer. One of the components of a health information system (HIS) is a prescribing assistance software (PAS) that helps to limit adverse drug events (ADEs). When physicians use software for healthcare, most of the data is entered manually into the PAS, which reduces the time spent on care. To overcome this barrier, we propose to provide a natural language interface to PAS so that practitioners can record their prescriptions using speech at the point of care using a smartphone. Such a system would allow practitioners to use a PAS on the go and allow them to get as close as possible to the most natural way of prescribing medications.

The general approach we have adopted is a task-oriented dialogue system coupled with a PAS destined for prescribers. One of the main challenges was to design a complete specialized dialogue system with processing mostly based on deep machine learning methods without available training data. To circumvent this problem, we proposed an iterative method coupling corpus collection, automatic text generation and expert modeling. We present a conversational modeling validated with medical experts and a detailed study of the characteristics of drug prescriptions from a NLP perspective. One of the main components of a dialogue system is focused on the natural language understanding (NLU) process, which is addressed by a slot-filling approach. Our method of data collection and generation allowed us to create a balanced corpus covering the whole prescription semantics we defined. This corpus allowed the initial learning of the NLU model as well as the boostrapping of the dialogue system.

To validate our approach and to collect realistic oral prescription data, we proposed an experiment with the dual objective of evaluating the system and collecting data. The evaluation, which included 55 people, including 34 medical experts, showed that the performance of NLU was satisfactory with an F-measure of 90% for a task success rate of more than 75% for the medical experts. These results are comparable to those obtained on the initial corpus, which confirms that the approach adopted during the thesis was valid. In order to promote reproducible research, the aligned oral corpus (speech-transcription-semantics) comprising more than 4 hours of recordings will be released to the community.

This thesis was carried out within the framework of a CIFRE (convention industrielle de formation par la recherche) collaboration between the company Calystene SA and Laboratoire Informatique de Grenoble (LIG).

Keywords : Spoken dialogue systems, natural language understanding (NLU), biomedical computing, biomedical NLP

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

Introduction

1.1 Motivation

Les Systèmes d'Information Hospitaliers (SIH) se sont imposés dans les établissements de santé pour améliorer l'organisation ainsi que la qualité et la traçabilité des soins (Lau et coll., 2010) qui nécessitent une maîtrise de toute la chaîne d'information en rapport direct ou indirect avec le patient. L'un des composants majeurs d'un SIH est le Logiciel d'Aide à la Prescription (LAP). Un LAP assiste le médecin pendant la saisie informatique des prescriptions et permet de tracer l'origine de celle-ci dans le parcours de soin du patient. Il existe aujourd'hui un consensus sur son efficacité dans la prévention des Événements Iatrogènes Médicamenteux (EIM) (Mille et coll., 2005). Les EIM, c'est-à-dire les erreurs liés à la prise de médicament, avaient fait en 2006, aux États-Unis, 1,5 million de victimes (Aspden et coll., 2006).

En France, selon une étude nationale, le nombre d'hospitalisations liées à un événement iatrogène grave était estimé entre 350 000 à 460 000 par an (Haury et Cases, 2005). Une autre étude réalisée sur les EIM confirme ces études américaines : selon une étude réalisée sur 5000 ordonnances, 1.7 % d'erreurs ont été détectées. Leur analyse détaillée montre que 42 % de ces erreurs correspondent aux mentions incomplètes indispensables à la préparation pharmaceutique. Sont également présentes des erreurs de surdosage (32 %) ou de sous-dosage (6 %) (Augry et coll., 1998). D'autres travaux menés dans le milieu hospitalier ont montré qu'un grand nombre d'erreurs entraînant des EIM sont évitables (Barker et coll., 2002; Kaushal et coll., 2001) et que les technologies de l'information et de la communication peuvent permettre de les réduire (Agrawal, 2009). Dans un logiciel intégré de gestion du circuit des médicaments et des soins, la prescription médicamenteuse doit obligatoirement être structurée ce qui la rend plus facile à vérifier et traiter. Par ailleurs, les prescriptions ont un un impact direct sur la gestion des stocks de la pharmacie et sur l'organisation du travail des services infirmiers dans la préparation et l'administration des traitements. Ceci explique la forte pénétration des LAP dans les établissements de santé afin de réduire le nombre d'erreurs de médicaments au cours du processus de prescription, de transcription et d'administration.

Cependant, avec l'évolution des systèmes d'information des établissements de santé, les professionnels de santé sont tenus de saisir de plus en plus d'informations numériques. Par exemple, les LAP imposent une saisie informatique contrôlée et détaillée des informations

de prescription qui pouvaient rester implicites dans le passé. Cette saisie laborieuse diminue le temps consacré aux soins et interrompt le clinicien dans ses tâches. Ainsi, il arrive régulièrement que certaines prescriptions ne soient pas entrées dans le système mais communiquées de manière manuscrite ou orale. Pour résoudre ce problème, nous proposons dans cette thèse, d'étudier et d'implanter un système dialogique interfacé avec les LAP qui permettrait aux praticiens d'enregistrer leurs prescriptions oralement en utilisant un langage plus naturel, plus proche de leur pratique habituelle.

Une interface dialogique orale et mobile a plusieurs avantages. Les logiciels complexes utilisés dans les différents centres de santé nécessitent une phase d'apprentissage qui peut être réduite en utilisant une interaction en langage naturel. De plus, les cliniciens peuvent utiliser leurs propres téléphones et se familiariser rapidement avec l'interface mobile tout en permettant une identification aisée sur le système. De telles interfaces permettraient également aux cliniciens de prescrire sur le lieu de soins, ce qui leur ferait gagner du temps et leur permettrait de se concentrer sur les soins (Altieri et coll., 2006). En outre, un système de prescription basé sur le dialogue pourrait signaler les médicaments qui ne sont pas disponibles dans la pharmacie d'officine et informer le prescripteur des effets indésirables potentiels des médicaments. Il pourrait aussi fournir des informations personnalisées liées aux patients (par exemple, sur les allergies) afin que le praticien puisse adapter la prescription en temps réel. En outre, la politique de dialogue pourrait inclure les choix relatifs au centre de soins concernant les médicaments et les traitements disponibles ainsi que les meilleures pratiques.

1.2 Partenaire Industriel

Cette thèse s'est déroulée dans le cadre d'une collaboration CIFRE entre le laboratoire d'informatique de Grenoble (LIG), au sein de l'équipe GETALP, et la société Calystene. Créé en 1992, Calystene développe des solutions informatiques médicales et administratives pour les hôpitaux et les cliniques. Calystene a une expertise reconnue en matière de dossier patient partagé, de prescription-dispensation nominative, de chaîne des soins et du circuit du médicament sécurisé ainsi que dans d'autres domaines en liaison avec les établissement de soins.

Notre travail s'intègre au logiciel Futura Smart Design© qui permet la gestion du circuit du médicament, la chaîne des soins, ainsi que des outils de planification, transmission et de facturation. Le module d'aide à la prescription de Futura SD© autorisé par la Haute Autorité de Santé (HAS) permet la génération des ordonnances médicamenteuses et non médicamenteuses tout en effectuant des contrôles automatisés pour renforcer la sécurité du processus de prescription.

Calystene a également un réseau de clients qui utilisent Futura SD© dont nous avons pu bénéficier grâce au contact avec des experts du domaine et des informaticiens spécialistes de systèmes informatiques de santé.

1.3 Positionnement

Dans cette thèse, nous nous plaçons dans un contexte d'interaction vocale sur terminal mobile qui permet de répondre à deux problèmes observés : l'interaction non naturelle avec les LAP et le manque d'accès immédiat à un système informatique connecté. La solution que l'on propose se rapproche de l'assistant personnel, de plus en plus présent dans le domaine de santé et qui serait accessible via un smartphone ou une tablette. En effet, il y a de plus en plus de systèmes de dialogue qui émergent dans le domaine de santé avec des objectifs divers (Bickmore et Giorgino, 2006; Kearns et coll., 2019) : automatisation des processus cliniques et monitorage (Black et coll., 2005; Adams et coll., 2014), formation du personnel médical (Llanos et coll., 2015; Talbot et coll., 2016), systèmes pour la santé mentale (Amini et coll., 2013; Fitzpatrick et coll., 2017), etc. Dans notre cas, le système de dialogue est destiné aux prescripteurs, c'est-à-dire toute personne ayant l'autorité de prescrire un traitement médicamenteux (p.ex. une personne maïeuticienne).

L'une des premières motivations d'un système vocal est liée à l'utilisation des dictaphones et l'interaction vocale qui est une pratique courante en médecine. Les applications mobiles de santé ainsi que les systèmes de reconnaissance automatique de la parole (RAP) ont vu un essor d'utilisation depuis l'intégration des SIS dans les hôpitaux, notamment après les avancés technologiques qui ont permis d'obtenir un taux d'erreur de mots (WER) d'environ 7-15 % en utilisant des logiciels commerciaux (Liu et coll., 2011). De plus, concernant les prescriptions médicales, le langage utilisé est très spécifique, ce qui rend la tâche d'analyse plus focalisée et réduit considérablement le vocabulaire à prendre en compte. En outre, l'interaction dialogique permet de recouvrir des erreurs et de s'adapter à l'utilisateur en respectant une interaction naturelle. Ensuite, la définition des connaissances et des stratégies discursives pourra grandement s'appuyer sur la connaissance métier intégrée dans les LAP.

Un autre aspect novateur de cette thèse réside dans son application aux prescriptions médicales **orales**. En effet, si les textes médicaux ont toujours fait partie des applications du TALN pour le diagnostic ou les pathologies (Xu et coll., 2010), les recherches sur l'extraction de prescriptions médicales sont plus rares. Par exemple, Tao et coll. (2018) propose un système d'extraction de prescriptions semi-supervisé à partir des données d'extraction d'information à partir des comptes rendus médicaux (Uzuner et coll., 2010b). Ces travaux sont souvent effectués sur des textes plus conséquents qu'une simple prescription tels que les documents de sortie d'hospitalisation (*discharge summaries*) ou les dossiers électroniques de patients.

Concernant la saisie de prescriptions à l'oral, nous avons trouvé quelques travaux existants. Le premier est FreePharma© cité dans un chapitre en 2006 (Dos Santos et coll., 2006). Il était décrit comme étant capable d'extraire des prescriptions médicales à partir de la parole dictée depuis un *Personal Digital Assistant* (PDA). Cependant, aucun détail technique n'est fourni et la référence renvoie vers un lien mort. Un autre travail connexe est le projet européen Mobi-Dev (Altieri et coll., 2006) qui vise à fournir la prochaine génération de dispositifs mobiles pour les cliniciens sur le lieu de soins. Sur le site Web de ce projet, il est indiqué que

les PDA dotés d'un système de reconnaissance automatique du langage naturel reliés aux systèmes d'information des hôpitaux permettent aux professionnels de santé d'atteindre un degré de mobilité sans précédent, dans un contexte hospitalier et au-delà, tout en réduisant les erreurs médicales. Cependant, malgré tous nos efforts, nous n'avons pas trouvé de publications scientifiques liées à ce projet. L'un des projets proche est celui de Ikhu-Omoregbe et Azeta (2010) proposé en 2010. Les auteurs présentent un système de dialogue oral permettant la saisie des prescriptions afin d'éviter les erreurs et potentiellement détecter les EIM et les interactions médicamenteuses. Le système de dialogue qu'ils proposent s'appuie sur VoiceXML (Lucas, 2000) où les informations sont dictées dans un ordre précis par le biais d'un appel téléphonique. Cependant, les auteurs ne présentent aucune évaluation du système ni de réalisation concrète.

Ces travaux nous montrent qu'il y a eu un intérêt avant les années 2010 pour une application de saisie de prescriptions orales pour les cliniciens dans un contexte hospitalier. Cependant, le peu de publications scientifiques sur le sujet et le manque de détails techniques ni d'évaluation dans une situation écologique indiquent que ces projets n'ont pas aboutis. Plus récemment, quelques applications de saisie de prescriptions orales, principalement émanant de chercheurs indiens, ont été proposées (Mahatpure et coll., 2019; Sanjeev et coll., 2021; Babu et coll., 2021). L'un de ces travaux (Mahatpure et coll., 2019) reste très concentré sur l'implantation logicielle d'une dictée vocale des prescriptions médicales et leurs transcriptions pour leurs transmissions afin d'éviter les incompréhensions liées à la lecture des ordonnances. Leur évaluation sur une population non-spécifiée montre que la saisie par dictée vocale d'une prescription serait plus rapide comparée à une saisie sur un clavier de smartphone. Aucune analyse sémantique ou d'erreur n'est mentionnée. Nous pouvons également citer un système de dialogue récent (*DocPal*) permettant d'accéder et de modifier les comptes rendus médicaux (EHR) à l'oral (Bhatt et coll., 2021). Le questionnaire de satisfaction sur ce premier prototype montre que les participants ont trouvé cet assistant vocal facile à utiliser, rapide et précis.

En France, les systèmes de dialogue en santé sont en plein essor tant du point de vue industriel qu'académique. Par exemple, Sanofi, dans son rapport [1] sur les systèmes de dialogue de santé publié en 2018 présente plusieurs domaines d'application des systèmes de dialogue, leurs enjeux sous la forme d'un cahier de retour d'expérience sur le développement d'un système de dialogue dans le domaine de la santé. Également, le *think tank* multidisciplinaires de la santé digitale : le lab e-santé a produit un livret blanc sur le sujet des systèmes conversationnels dans le domaine médical [2]. Tous ces travaux de réflexion montrent qu'il y a de plus en plus d'intérêt sur le sujet et suscitent l'attention de l'industrie.

Malgré l'évolution rapide du domaine, les travaux sur le traitement automatique des prescriptions médicamenteuses restent rares. Il existe des travaux qui focalisent sur la reconnaissance optique de caractères (OCR) permettant de numériser les ordonnances ma-

1. https://www.sanofi.fr/fr/-/media/Project/One-Sanofi-Web/Websites/Europe/Sanofi-FR/Newsroom/nos-publications/Livre-blanc-BOT-V03_BD.pdf
2. https://www.ticsante.com/documents/201907041716270.Livre_blanc_chatbot_du_Lab-esante.pdf

nuscrites (Kim et coll., 2015; Sarzynski et coll., 2017), classification automatique des prescriptions (Carchiolo et coll., 2019) ou d'autres applications de TALN sur les prescriptions. Cependant, nous n'avons pas trouvé de travaux qui visent à saisir des prescription médicale via un système de dialogue complet à l'oral interagissant avec un LAP.

1.4 Objectifs du travail de thèse

L'objectif principal de cette thèse est de définir comment les techniques du dialogue oral à l'état de l'art peuvent être couplées à un système LAP métier pour permettre la saisie de prescriptions médicales sur le lieu de soins en utilisant un terminal mobile en langage naturel.

Nous abordons le problème dans le cadre du TALN permettant d'obtenir une structure de données à partir des énoncés oraux. Le choix d'une interaction dialogique où l'énoncé oral est analysé puis désambiguïsé/complété/corrigé à travers des tours de parole permet au prescripteur de garder la maîtrise de la prescription.

Le défi principal à relever dans cette thèse était le manque de ressources pour l'analyse, la conception et l'évaluation. Il n'existait en effet aucun exemple de système de prescription orale ni aucune donnée de prescriptions numériques accessibles en français. Un autre défi, plus conjoncturel, était l'accès au terrain médical en pleine période de pandémie.

Concernant la modélisation de la solution, nous nous sommes orientés vers une saisie des prescriptions à l'oral en adoptant l'approche *slot-filling* (remplissage d'attributs) qui permet d'obtenir des informations compatibles avec le système d'information du LAP. La sémantique des prescriptions médicamenteuses devait être conçue pour, d'une part, correspondre à la sémantique métier, et d'autre part, pour faciliter le flux dialogique lors de l'échange entre les différents acteurs. L'une des contributions de cette thèse est d'avoir établi cette sémantique en tirant profit des modélisations métier existantes et en interagissant avec des experts des prescriptions médicamenteuses pour d'affiner cette modélisation et valider l'approche.

Les modèles à l'état de l'art de traitement du dialogue reposent sur l'apprentissage automatique supervisé qui lui-même nécessite un ensemble de données d'entraînement conséquent. Or, cet ensemble de données était inexistant pour le domaine de la thèse. Il était d'ailleurs impossible de le collecter sur le terrain, car l'application était également inexistante. Enfin, une collecte en magicien d'Oz ne pouvait pas être considérée étant donné le manque de disponibilité des experts. Nous avons donc mis au point une démarche itérative permettant d'initialiser des modèles de traitement initiaux (NLU, politique de dialogue) pour collecter des données d'interaction avec des volontaires puis confronter les modèles aux experts afin de valider le processus. En utilisant cette démarche, il a été possible de mettre au point un système complet dialogique à l'état de l'art ainsi que récolter un corpus oral de prescriptions médicamenteuses annoté pour la compréhension et le dialogue destiné à être diffusé à la communauté. Ce système a pu être validé avec des experts médicaux et entièrement connecté à un LAP professionnel.

1.5 Plan du manuscrit

Le manuscrit est divisé en six parties. Le chapitre 1 correspond à l'introduction de ce travail. Y sont présentés la motivation et le partenariat qui ont mené à ce manuscrit, ainsi que les objectifs du travail de thèse.

Le chapitre 2 développe l'état de l'art des trois domaines auxquels appartient cette thèse : TAL biomédical, les systèmes de dialogue et la compréhension du langage naturel. Dans la partie dédiée au TAL biomédical, ses caractéristiques et la représentation de connaissances utilisées dans le domaine de la santé sont expliquées. Ensuite, dans la deuxième partie, nous présentons les caractéristiques des conversations humaines ainsi que les systèmes de dialogue, notamment ceux utilisés dans le domaine de la santé. Enfin, nous abordons les systèmes de compréhension du langage utilisé dans le domaine général ou également dans le biomédical.

Le chapitre 3 présente notre problématique, nos questions de recherche et la méthode suivie tout au long de cette thèse.

Le chapitre 4 détaille les modèles de compréhension automatique du langage que nous avons utilisés dans cette thèse ainsi qu'une évaluation de ces systèmes sur un corpus de prescriptions médicamenteuses acquis avec des données augmentées.

Le chapitre 5 aborde notre démarche du point de vue de la modélisation du dialogue et présente une première évaluation sur le système de dialogue sur une pré-expérimentation.

Dernièrement, le chapitre 6 décrit l'expérimentation à moyenne échelle que nous avons mise en place pour l'évaluation du système de dialogue et les modèles de compréhension. Cette expérimentation a permis de recueillir des données réalistes pour évaluer a posteriori nos modèles et pour distribuer cette ressource à la communauté scientifique.

CHAPITRE 2

État de l'art

Notre travail s'inscrit dans le cadre d'un champ de recherche pluridisciplinaire qui rassemble le domaine de la linguistique, de l'informatique et de la médecine : celui du traitement automatique des langues (TAL) biomédicales. Ce dernier englobe de nombreuses thématiques de recherche telles que l'extraction d'information à partir des textes biomédicaux, le dialogue en langage naturel, les plongements de mots et de concepts spécialisés, etc. Dans ce chapitre, nous allons premièrement faire un état des lieux des travaux relevant du domaine du TAL biomédical en 2.1. Nous décrirons ensuite les caractéristiques de l'une des langues de spécialité, plus particulièrement celle utilisée par les experts du domaine biomédical. Nous présenterons brièvement les thésaurus, ontologies et codages médicaux utilisés dans le domaine.

La conception d'un système de dialogue nécessite la définition claire d'un domaine et d'un espace sémantique explicite, notamment dans le cadre d'un système de dialogue orienté but. Cette modélisation dépend largement des concepts provenant des caractéristiques des conversations humaines. Dans la partie 2.2, nous allons aborder les actes de paroles et les théories de la communication employés dans un échange conversationnel. Cette section est suivie de 2.3 où sont abordés les systèmes de dialogue orientés but, en mettant l'accent sur ceux utilisés dans le domaine de la santé. Après une présentation des différentes approches à la conception des systèmes de dialogue, nous présenterons en détail les composants d'un système de dialogue oral à base de cadres dans 2.4.

Pour un système de dialogue modulaire, l'un des composants les plus importants est la compréhension automatique du langage. Les erreurs qui surviennent dans la reconnaissance automatique de la parole (RAP) et la compréhension du langage naturel (NLU) peuvent avoir des effets à cascade dans les étapes suivantes de la chaîne de traitement (pipeline), notamment dans le suivi de l'état du dialogue. La section 2.5 va présenter les approches utilisées dans le domaine de compréhension automatique du langage naturel.

Les approches récentes en TAL s'appuient sur l'apprentissage sur de grandes quantités de données (annotées ou non). Dans le domaine du TAL biomédical, malgré les défis, tâches collectives (*shared tasks*), et autres initiatives, l'accès aux données biomédicales reste difficile notamment dans une langue autre que l'anglais. Dans la partie 2.6, nous allons présenter les entrepôts de données, les tâches partagées, et les défis autour de la thématique d'extraction d'information biomédicale. Cet état de l'art se conclura par la partie 2.7 en présentant quelques techniques utilisées dans le domaine biomédical face à ce défi qu'est le manque

de ressources.

2.1 Domaine du TAL biomédical

Le TAL biomédical fait parti du grand domaine qu'est le traitement automatique des langues naturelles qui s'intéresse à traiter automatiquement les énoncés en langue naturelle et à leur contenu dans un contexte médical (Zweigenbaum, 1997). Dans ses anciens jours, la préoccupation majeure de ce domaine pluridisciplinaire portait sur la constitution de lexiques médicaux (McCray et coll., 1994; Burgun et coll., 2003), la création de terminologies (Spackman et coll., 1997; Bodenreider, 2004), la formalisation (Lipscomb, 2000) et la représentation des connaissances (Zweigenbaum et coll., 1995) dans le domaine médical. L'étude et le développement de ces différentes thématiques portent majoritairement sur des textes médicaux : comptes-rendus médicaux, notices de médicaments, articles scientifiques dans le domaine biomédical, etc. Malgré l'informatisation et la structuration des informations médicales, de nombreuses données cliniques biomédicales restent sous forme de texte libre dans des publications scientifiques, des dossiers cliniques, des comptes-rendus, etc (Kreimeyer et coll., 2017). L'exploitation de ces textes d'une manière automatique et formelle a permis la mise en place de protocoles et applications qui ont été jugés comme bénéfiques dans beaucoup de contextes, notamment pour le développement de la médecine, l'aide à la décision automatique dans un contexte clinique, le diagnostic préventif, etc. (Wang et coll., 2018).

En revanche, le domaine du TAL biomédical n'est pas simplement du TAL appliqué dans le domaine de la santé mais plutôt un domaine qui s'inscrit dans les problématiques du domaine de l'informatique médicale avec des techniques de TAL Zweigenbaum (1997). De nos jours, l'évolution du TAL ainsi que l'informatique médicale, l'accessibilité de ressources et la disponibilité des données ont permis des avancées impressionnantes Cambria et White (2014); Young et coll. (2018). Cette même tendance a été appliquée et constatée dans le domaine du TAL biomédical. La revue systématique de Wu et coll. (2020) montre que la recherche impliquant l'apprentissage profond a été plus que doublée chaque année à partir 2018 et montre que dans certaines tâches, l'utilisation des méthodes à base de réseaux de neurones est prévalente, par exemple concernant l'extraction d'information ou la reconnaissance d'entités nommées (92% des papiers) Wu et coll. (2020). Cette tendance n'est pas surprenante car la création de règles ou l'ingénierie des caractéristiques (*feature engineering*) sont très coûteuses et chronophages, la participation des experts médicaux étant souvent nécessaire. Pour cette même raison, nous avons choisi de procéder avec des méthodes d'inférence pour la gestion du dialogue et la compréhension.

Dans cette section de l'état de l'art, les caractéristiques de la langue employée dans le domaine biomédical (à l'oral et à l'écrit) vont être abordées ainsi que la description des méthodes, codages et efforts utilisés dans la normalisation des termes biomédicaux.

2.1.1 Caractéristiques de la langue biomédicale

La conception théorique d'une interaction langagière dans le domaine biomédical et les traitements automatiques qui seront réalisés nécessitent l'étude des caractéristiques du domaine. *La langue biomédicale* pourrait être qualifiée comme une langue de spécialité basée sur une langue de référence (français dans notre cas), pour rendre compte de connaissances spécialisées dans le domaine de la biologie et du médical. Lorsqu'on parle d'une langue de spécialité, il ne s'agit pas d'une langue à part entière, mais plutôt d'une langue dans laquelle on trouve des tournures syntaxiques et des expressions qui n'auraient pas existé dans la langue de référence (Charnock, 1999). Il ne s'agit pas non plus uniquement de transmission de connaissances des sujets techniques par une terminologie médicale, mais aussi de la présence d'un vocabulaire biomédical, de la fréquence d'occurrence d'unités terminologiques et de l'énonciation d'un discours qui relèvent de styles ou de modes d'expression très variés, allant du registre soutenu au familier (Charnock, 1999).

Cette langue de spécialité comporte un caractère d'efficacité : cela permet aux praticiens de santé de condenser en un nombre réduit de mots une grande quantité d'informations (Vecchiato et Gerolimich, 2013). En effet, sur le plan sémantique, les phrases sont longues et comportent souvent plusieurs constructions, voire même des constructions non finies comme l'exemple ci-dessous tirées de l'article de Vecchiato et Gerolimich (2013) :

« Névralgie d'Arnold (conflit du nerf occipital avec la charnière osseuse) : douleur en éclair, déclenchée par les mouvements du cou, partant de la charnière cervico-occipitale et irradiant en hémicranie jusqu'à la région frontale » (Vecchiato et Gerolimich, 2013)

Au niveau lexical, l'utilisation de mots d'emprunts est très courante tel que nuque (Arabe), alvéole (latin), stroke-center (anglais). La médecine est un domaine de recherche impliquant la création de nouveaux termes, en particulier empruntés de l'anglais (blind test, scanning, etc.). L'utilisation de termes non codifiés créée des risques et engendre des malentendus. C'est l'une des raisons pour lesquelles la projection d'expressions en langue naturelle sur la terminologie de référence a été la motivation la plus répandue dans la naissance du TAL biomédical (Zweigenbaum, 1997). Au niveau morphologique, la terminologie biomédicale comporte beaucoup de mots composés dont les bases sont souvent empruntées du latin et du grec (Kallel, 1999). La plupart des mots complexes sont de nature compositionnelle. Pour cette raison, d'un point de vue du traitement automatique, faire revenir les formes fléchies à leurs formes de base contribue beaucoup aux performances des systèmes, notamment dans la recherche d'information et l'indexation (Zweigenbaum et coll., 2001).

De plus, les professionnels de santé utilisent des termes spécifiques même lorsqu'il y a des équivalents en langue naturelle en privilégiant une communication fonctionnelle et efficace. Par exemple, pour le terme « gêne respiratoire », il est possible d'utiliser un équivalent comme une difficulté à respirer. Cependant, comme dans tous les domaines techniques, l'utilisation d'un lexique spécifique vise la précision et à ne pas engendrer de confusion entre spécialistes.

Même si ces caractéristiques soulignent des problèmes de traitement du texte, elles sont

tout de même applicables aux énoncés oraux. En effet les noms d'emprunt, l'utilisation très fréquente des abréviations rend le traitement automatique de la langue qu'elle soit textuelle ou orale plus délicat. A ceci, s'ajoute un phénomène particulièrement difficile à traiter : les médicaments à présentation et à consonance semblables (*look-alike sound-alike drugs* en anglais ou *LASA*). Ce sont des médicaments dont le nom ou l'apparence peuvent prêter à confusion, constituant ainsi un enjeu important pour la santé publique. Par exemple, en Suisse, Kundig (2011) reporte que 7% des erreurs de délivrance (parfois fatales) sont liés aux médicaments LASA et, malgré les efforts de les réduire, Thompson (2008) rapporte l'existence de milliers de médicaments facilement confondus aux États-Unis. Les ordonnances manuscrites, notamment celles comportant une écriture difficilement lisible contribuent à ces confusions (Kundig, 2011).

Ce phénomène est doublement intéressant dans la conception d'un système oral sur les prescriptions médicamenteuses : A) la reconnaissance de la parole sujette à des erreurs de confusion entre les médicaments à consonance semblable et B) les confusions visuelles de graphèmes à l'écrit des noms de médicaments. Les systèmes hospitaliers et les LAP emploient diverses stratégies afin de limiter ces erreurs, tels que des codes couleurs ou plusieurs alertes de vérifications afin de ne pas engendrer des confusions entre les médicaments LASA. Le tableau 2.1 montre quelques exemples de médicaments à consonance semblables.

Médicaments LASA
Metfin©, Metformine©,
Avalox©, Avonex©
Penicillamine, Penniciline
Retrovir©, Norvir©
Cyclosérine, Ciclosporine
Sulfadiazine, Sulfasalazine
Rifampin, Rifaximin
Valganciclovir, Valacyclovir
Metolazone, Métoprolol, Metoclopramide
Loratadine, Lovasatatine

TABLE 2.1 – Exemple de médicaments à consonance semblables rapportés dans (Kundig, 2011; Lambert et coll., 2019)

L'extrait des exemples présentés sur le tableau 2.1 montre que les confusions ne concernent pas que les spécialités (noms commerciaux des médicaments) mais aussi les dénominations communes internationales (DCI). Afin d'éviter ces erreurs, différentes approches automatiques sont proposées. Par exemple Lambert et coll. (2019) propose une mesure de similarité et évalue son approche sur un échantillon de 506 commandes de médicaments au sein d'un établissement clinique. Les IHM parfois intègrent des stratégies pour attirer l'attention sur l'écriture d'un médicament LASA pour éviter les erreurs.

2.1.2 Représentation des connaissances biomédicales pour la prescription médicamenteuse

Le traitement automatique du texte ou de la parole nécessite une transformation du contenu langagier vers une représentation de connaissances exploitable par une machine. Cette représentation dépend largement de la modélisation qui doit être faite au préalable permettant de représenter les informations pertinentes que le système informatique pourrait exploiter (Zweigenbaum, 1999). Ces modèles représentent les termes du domaine ainsi que leurs relations du plus général au plus spécifique sous forme de terminologie ou d'ontologie. Le niveau de formalisation d'une ressource lexicale dépend de son utilisation et sa conception, il s'agit donc des ressources *termino-ontologiques* qui peuvent aller d'un thésaurus à une ontologie (Névéol, 2018).

La première ressource lexicale officielle publiée par *National Library of Medecine* (NLM) est le *MeSH*(Medical Subject Headings en anglais) introduite en 1960 (Lipscomb, 2000). Introduit pour indexer des articles de la littérature à l'aide des titres des sujets médicaux (d'où vient son nom), ce thésaurus du domaine médical est revu tous les ans. MeSH est aussi traduit dans d'autres langues que l'anglais. Par exemple, la version française est traduite et publiée par l'INSERM (Institut National de la Santé Et de la Recherche Médicale). Une autre ressource importante et historique dans le domaine est la classification CIM-10 (Classification Internationale des Maladies)(OMS, 1993). Cette classification est associée à un système de codage permettant de classifier des maladies, des symptômes, des troubles mentaux, etc. vers des codes internationaux permettant le codage en morbi-mortalité proposé par l'OMS (Organisation Mondiale de la Santé) [1]. D'autres terminologies de référence sont les terminologies MedDRA (dictionnaire médical des activités de réglementation) (Brown et coll., 1999) utilisées notamment dans le domaine de pharmacovigilance et le Snomed International (Nomenclature systématique de médecine) (Côté et coll., 1993) qui est une nomenclature destinée à l'encodage médical plus fin des dossiers électroniques des patients. La nomenclature Snomed comporte des relations transversales plus complexes entre les concepts qui sont reliés par l'hyponymie et la méronymie et se rapproche plus vers une ontologie (Merabti, 2010).

Ces terminologies historiques citées ci-dessus, toujours utilisées et mises à jour en 2021, font partie de l'UMLS (Unified Medical Language System) développé par le NLM et comportent des millions de termes du domaine biomédical (Bodenreider, 2004). L'UMLS regroupe les ressources termino-ontologiques et établit des équivalences entre les concepts dénotés par ces différentes ressources (Névéol, 2018). Pour extraire des concepts et les relier avec leurs correspondances en UMLS, différents outils existent : le plus historique Metamap (Aronson, 2001), MedLee Friedman (2000) ou cTakes (Savova et coll., 2010). Ces outils sont des systèmes à base de règles et effectuent de la recherche de motifs associés à des ressources externes telles que le thésaurus multilingue UMLS. Les approches récentes se

1. https://www.cepidc.inserm.fr/causes-medicales-de-deces/classification-internationale-des-maladies-cim

focalisent sur l'apprentissage automatique, mais nécessitent des données d'entraînement annotées (Fraser et coll., 2019).

Le réseau sémantique d'UMLS est assez large, et les concepts utilisés dans les traitements et les corpus dépendent des domaines et des tâches. Pour représenter les prescriptions orales, la ressource terminologique concernant les médicaments et leurs interactions est le RxNorm (Liu et coll., 2005). Dans RxNorm, un médicament est décrit à l'aide de 10 concepts comme les ingrédients, la composition chimique, la forme du dosage, la puissance, etc. Cependant, le niveau de détail sur la composition chimique est fait pour prendre en compte les calculs réalisés dans les LAP mais ne prennent pas en compte la réalité d'une prescription formulée à l'oral, notamment en français où il peut y avoir des formulations complexes liées à la durée et la fréquence de la prescription. La ressource la plus proche permettant de représenter les informations sur une prescription médicale est l'ontologie PDRO proposée par Ethier et coll. (2018). Comme le RxNorm, cette ontologie est conçue pour relier les concepts d'une prescription électronique aux concepts normalisés proposés par les auteurs. Dans cette ontologie, les conditions et les durées de prescription sont modélisées en précisant les motifs de début, de continuation et de fin. La figure 2.1 présente un exemple de prescription avec les concepts de l'ontologie PDRO présenté par Ethier et coll. (2018).

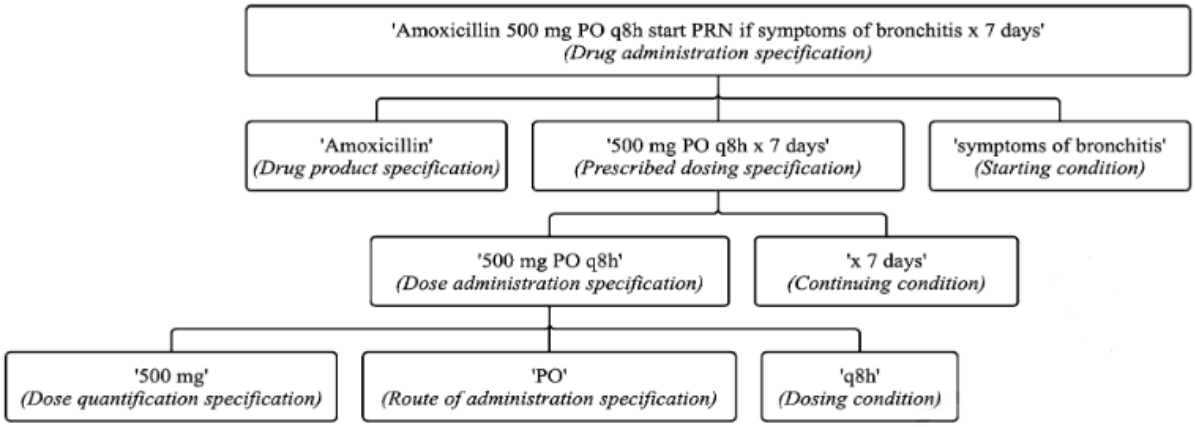


FIGURE 2.1 – Extraction de concepts à partir d'un exemple de prescription en anglais par l'ontologie PDRO (Ethier et coll., 2018)

Sur la Figure 2.1, les auteurs montrent l'extraction des concepts selon l'ontologie PDRO. Cette modélisation ne comporte pas autant de détails sur la composition même du médicament, mais distingue bien la dose du médicament (drug product specification) et le dosage indiqué dans la posologie (prescribed dosing specification). Contrairement à l'ontologie RxNorm, les concepts définissant la quantité et l'unité de mesure ne sont pas explicitement dissociés. Cette notion est importante dans la conception de l'interaction conversationnelle, car la politique du dialogue doit prendre en compte toutes les spécificités des concepts et doit distinguer la valeur numérique de l'unité de mesure (500,mg) ou de la durée(7,jours).

Cette modélisation prend en compte les réalités des comptes rendus hospitaliers où les praticiens expriment les conditions de début et justifient leur prescription par rapport au diagnostic établi, etc. En revanche, une ordonnance de sortie destinée au pharmacien

et au patient est soumise à des réglementations spécifiques et ne pourra pas être modélisée comme précédemment. Ces ressources permettant de décrire une prescription électronique se concentrent sur la composition des médicaments et leur appartenance à des classes pharmaco-thérapeutiques.

Pour représenter les connaissances issues du domaine des prescriptions médicamenteuses et l'adapter aux habitudes de prescription en français, nous nous sommes basés sur la ressource terminologique RxNorm et l'ontologie PDRO. Cependant, ces terminologie pensées pour des prescriptions à l'américaine où notamment les fréquences et les rythmes sont exprimées par des expressions latines (qhs, qds, qid, etc.). Or, lorsqu'on évoque les habitudes de prescription, il faut préciser qu'il y a une culture de la prescription qui peut varier en fonction des praticiens, leurs spécialités ou leurs institutions. Ainsi, la prescription médicamenteuse n'a pas le même contenu langagier selon les pays et ne comporte pas les mêmes obligations juridiques.

Pour la prescription en France, nous nous sommes appuyés sur la base de données médicamenteuses Thésorimed© qui utilise les systèmes de codage comme MeSH, CIM-10, classification ATC, etc. Pour le codage des médicaments, nous nous sommes référés aux codes des unités communes de dispensation (UCD) délivrés en établissements de santé proposés par l'Assurance Maladie et les codes CIP-13 visant à identifier chaque médicament remboursable dans la base de données nationale. Il est important de préciser que ce système de codage des médicaments n'est pas universel et seul un autre système : les dénominations communes internationales (DCI) est universel. Selon le Ministère des Solidarités et de la Santé [2], la prescription en DCI est obligatoire pour les médicaments génériques depuis 1er Janvier 2015. En pratique, l'objectif est de passer progressivement à des prescriptions en DCI après une phase de transition qui permettrait une identification universelle des médicaments. Un système de dialogue tel que celui envisagé dans cette thèse pourrait jouer un rôle de facilitateur dans le passage en DCI. Concernant les autres concepts définissant les prescriptions médicamenteuses, nous avons proposé une taxonomie adaptée ayant plus de concepts adaptés pour des aux prescriptions médicamenteuses et qui sera présentée en section 3.3.1.

2.2 Caractéristiques des conversations humaines

Les systèmes de dialogue utilisent des modèles de représentations qui sont inspirés directement de certaines théories du discours. Nous allons présenter dans cette section les théories les plus en lien avec nos travaux.

Les conversations orales entre les humaines sont beaucoup moins structurées qu'à l'écrit, plus courtes que les phrases rédigées et réfléchies et comportent des ellipses et des déictiques spatio-temporels implicites (Fouquet, 2004). Les théories du discours et les ana-

2. https://solidarites-sante.gouv.fr/soins-et-maladies/medicaments/professionnels-de-sante/prescription-et-dispensation/article/prescription-en-denomination-commune-internationale-dci

lyses sur les conversations humaines ressortent beaucoup de phénomènes linguistiques et extra-linguistiques qui régissent lorsque deux ou plusieurs humains entament une conversation. Dans un contexte purement conversationnel, il est très difficile d'analyser ou encore plus de modéliser les motivations intrinsèques des interlocuteurs. En revanche, nous nous situons dans un domaine de conversation particulier : le domaine d'une conversation dans le but d'effectuer une démarche, dans l'occurrence afin d'effectuer une prescription médicale.

De même, est-ce qu'une interaction humaine-humaine comporterait les mêmes caractéristiques d'une interaction humaine-machine ? Le constat de Fraser et Gilbert (1991) résume bien cette situation :

> « Le concepteur (designer) est pris dans un cercle vicieux : il est nécessaire de connaître les caractéristiques des dialogues entre les personnes et les automates pour pouvoir construire le système, mais il est impossible de savoir à quoi ressembleraient ces dialogues tant que le système n'a pas été construit. » (Fraser et Gilbert, 1991)

Dans ce contexte, nous allons aborder le côté performatif du discours et aborder les théories d'Austin et Searle et ainsi définir la notion des actes de paroles. Comme nous avons évoqué dans 2.1.1, la langue biomédicale privilégie une communication succincte, fonctionnelle et efficace. Nous allons faire le lien avec les caractéristiques de la langue biomédicale et les lois de communication définies par Grice. Par la suite, nous allons évoquer les concepts de la prise d'initiative dans le discours, les phénomènes d'inférence et l'implicature conversationnelle liée aux maximes de Grice.

2.2.1 Actes de parole

Les systèmes de dialogue sont fondés sur les principes des théories du discours, notamment sur la théorie des actes de langage, élaborée par John Austin (Austin, 1975). Cette théorie part du principe que la fonction du langage n'est pas seulement de décrire le monde, mais aussi d'accomplir des actions. Ce principe s'opposant à la conception descriptive du langage décrit la situation d'énonciation comme un acte social.

Dans ce contexte, un acte de parole peut être défini comme une intention communicative associée à un message linguistique. Lorsque deux humains interagissent, l'énonciateur produit un énoncé dans une langue avec pour objectif dont le locuteur, à partir de l'énoncé, du contexte communicatif et de la nature de la communication, infère l'acte de parole (Bach et Harnish, 1979). L'acte locutoire est la forme de surface, c'est-à-dire le sens de l'énoncé représenté par la grammaire. Quant à lui, l'acte illocutoire représente l'intention de l'énonciateur de susciter une attitude particulière chez le locuteur. Ces concepts qui constituent la partie primordiale des actes de discours ont été influencés par les théories d'Austin et de Searle. Les constituants des actes de paroles peuvent être résumés comme ainsi selon Austin :

— Acte énonciatif : L'énonciateur (S) produit un énoncé (e) dans une langue (L) destinée à un entendeur (H) dans un contexte communicatif (C)

— Acte locutoire : Le fait de produire une suite de sons ayant un sens dans une langue L.

— Acte illocutoire : L'énonciation de l'énoncé (e) transforme le rapport entre les interlocuteurs (intention rhétorique du locuteur)

— Acte perlocutoire : L'énoncé (e) provoque des effets dans la situation de communication (C) qui va au-delà du cadre linguistique et convoite le pouvoir de changer l'état du monde.

Au lieu d'étudier le discours dans son ensemble, les premiers travaux ont consisté en l'étude des énoncés de façon isolée en partant de ce même principe : les énonciateurs accomplissent des actes illocutoires tout au long de la conversation. Plus tard, Austin utilise aussi le terme *acte de dialogue* pour désigner également les actes de paroles dans un contexte dialogique. On trouve ce principe dans la conception des systèmes de dialogue où chaque énoncé est une sorte d'action réalisée par l'énonciateur (Jurafsky, 2000). L'équation 2.2.1 représente une formalisation de cette notion :

$$ActeDialogue = \underbrace{(a = x, b = y, ...)}_{attributs, valeurs} \tag{2.1}$$

Dans ce formalisme, l'acte de dialogue est représenté par l'intention du locuteur extraite à partir de l'énoncé avec un ensemble optionnel d'attributs valeurs qui correspondent aux cadres sémantiques. La taxonomie des cinq classes majeures d'acte de langage selon Searle, Austin et Bach et Harnish est présentée dans le tableau 2.2 :

Austin (1975)	Searle (1976)	Bach & Harnish (1979)	Définition
Expositifs	Assertifs	Assertifs	Affirmation d'un état de fait
Exercitifs	Directifs	Directifs	L'incitation de l'interlocuteur à agir
Promissifs	Promissifs	Promissifs	Engager le locuteur à une suite d'actions déterminées
Comportementaux	Expressifs	-	Exprimer un état psychologique
Verdictifs	Déclaratifs	-	Prononcer un jugement

TABLE 2.2 – Taxonomie des classes majeures d'actes de parole selon Austin, Searle et Bach & Harnish

Dans la modélisation de l'interaction conversationnelle d'un système de dialogue, nous nous appuyons sur des actes de parole qui peuvent être déclinés par la taxonomie décrite dans le tableau 2.2. Par exemple, dans le contexte médical, lorsque le prescripteur affirme une prescription médicamenteuse, cette dernière a un pouvoir déclaratif. Nous avons donc un énoncé performatif provenant d'un locuteur possédant un statut et une légitimité (diplôme d'Etat de docteur en médecine) suffisamment importants pour transmuter la parole en un acte effectif.

Cependant, le dialogue n'est pas un ensemble d'actes de paroles isolés, mais un type de communication entre deux personnes ou un groupe de personnes. En partant du principe qu'il y a une présomption communicative dans un dialogue, c'est-à-dire qu'il y a un partage des connaissances suffisantes entre les interlocuteurs permettant la communication (Bach et Harnish, 1979), l'énonciateur et le locuteur ont besoin de s'assurer qu'il y a une inter-compréhension dans le flux de dialogue.

Malgré les différences d'interprétation, le but communicatif fait adopter des stratégies dédiées à la convergence, notamment par un signe de tête. Les actions dédiées à montrer ce qui est compris par les locuteurs peuvent être dénommées « terrain d'entente » (common ground) (Stalnaker, 1978). Dans un système de dialogue textuel ou oral, la communication non verbale n'entrant pas en jeu, la manifestation de la compréhension est souvent faite par des actes langagiers dits *grounding* ou processus d'ancrage. En effet, selon Traum et Hinkelman (1992), l'un des caractéristiques les plus importantes des conversations repose sur ces signes d'accord et d'acquittement notamment dans le cadre d'un dialogue conversationnel.

2.2.2 Théories de la communication

La conception théorique de l'interaction conversationnelle d'un système de dialogue est étroitement liée aux fondements théoriques des actes de parole que nous avons résumés dans la partie 2.2.1. Notamment la notion abordée dans la théorie du discours ainsi que dans la formalisation de cette théorie, l'inférence de l'intention du locuteur. En effet, comme Searle et Bach&Harnish, Paul Grice contribue aux notions de pragmatique théorique et de communication « intentionnelle ». Les réflexions et les contributions majeures de Paul Grice portent sur l'analyse et la compréhension de l'intention réelle du locuteur dans une situation de communication. Pour décrire ce phénomène, il distingue le sens naturel de l'énonciation (relation entre le signifié et le signifiant) et le sens non naturel (pas de relation causale entre le signifiant et signifié). La signification de ce qui est dit est donc interprétable par l'effet que veut produire un énonciateur chez son interlocuteur (une croyance ou une volonté d'agir)(Grice, 1957). D'après (Grice, 1975), l'intention de départ du locuteur ne pourrait être définie que lorsque cette intention aura provoqué un certain effet sur son interlocuteur grâce à la reconnaissance de cette intention.

(A) Doliprane 1000 mg à diluer dans un verre d'eau, matin et soir
(B) Pendant combien de temps ?

Dans l'exemple ci-dessus, l'énonciateur attend que son interlocuteur fasse des inférences afin d'accéder au sens extra-linguistique porté par l'énoncé même. Dans l'exemple (A) le prescripteur nous informe sur le médicament, son dosage, avec une mention d'administration et le rythme de la prescription. Par contre, la voie d'administration et la forme galénique du médicament ne sont pas précisées explicitement dans l'exemple. La mention d'administration «à diluer dans un verre d'eau » permet cependant à l'interlocuteur d'inférer qu'il s'agit d'un comprimé effervescent et la voie d'administration comme la voie orale. Grice dénote cette notion comme **l'implicature conversationnelle**.

Selon Grice et coll. (1975), la conversation humaine suite une logique de « principe de coopération » : l'existence d'un but commun partagé par les participants d'un dialogue. De cette manière, dans une situation de communication, l'énonciateur peut s'attendre à ce que son/ses interlocuteurs suivent ce principe de coopérativité. Paul Grice détaille sa théorie et introduit la notion de maximes conversationnelles ou lois du discours dans la théorie pragmatique en 9 maximes classé en 4 catégories.

Catégorie	Règles
Maximes de quantité	Que votre contribution soit aussi informative que nécessaire Que votre contribution ne soit pas plus informative que nécessaire
Maximes de qualité	N'affirmez pas ce que vous croyez est faux N'affirmez pas la véracité d'un propos pour lequel vous manquez de preuves
Maxime de relation	Soyez pertinents
Maximes de manière	Evitez de vous exprimer de manière obscure Evitez l'ambiguité Soyez bref Soyez ordonné

TABLE 2.3 – Maximes conversationnelles de Grice (Grice, 1975)

Les maximes conversationnelles de Grice sont présentées dans le tableau 2.3. Dans la théorie du discours, Grice souligne l'importance de l'implicature conversationnelle et le définit comme étant les conséquences d'application ou de non-application de ces maximes dans une situation de communication. Le principe d'implicature est donc basé sur l'usage du langage est prend en compte les non-dits, les ellipses,etc. (Caelen, 2003).

Le principe de coopération ainsi que les maximes conversationnelles de Grice sont importants dans la conception des systèmes de dialogue, notamment pour les systèmes orientés but. Par exemple, Dybkjær et coll. (1996) proposent une extension aux maximes conversationnelles de Grice pour les systèmes de dialogue orientés but. Ces maximes peuvent être également utilisées dans le cadre de l'évaluation des dialogues produits entre le système et les utilisateurs (Jwalapuram, 2017), et a fondé les premières approches visant à extraire automatiquement l'intention du locuteur (Allen et Perrault, 1980).

2.2.3 Caractéristiques des conversations

Nous nous sommes intéressés à l'analyse du discours, aux actes de parole, aux intentions et aux implicatures conversationnelles afin de comprendre les caractéristiques des dialogues humains à partir des énoncés et la situation de communication. Or, les conversations humaines comportent d'autres caractéristiques liées à leurs structures et à la nature des conversations qui sont étroitement liées avec le principe de coopération de Grice ou les actes de paroles. Le domaine pluridisciplinaire de l'analyse conversationnelle étudie les mécanismes et l'organisation des conversations humaines.

La définition que nous avons retenue concernant l'analyse des conversations est celle de Vincent (2001) : « Dans une conversation, tout est un indice : de tension, d'indifférence, de

plaisir, d'harmonie, de demande de poursuite ou de volonté de tout interrompre » ; analyse basée sur le principe de décomposition des éléments du dialogue en petites unités afin d'en analyser les enjeux (Sacks, 1992). La décomposition en petites unités ou l'analyse des détails (indices) permettent d'étudier les phénomènes sociaux, l'interaction en elle-même et le déroulement de l'activité. Les phénomènes observés dans la structure des dialogues humains sont utilisés dans la conception et l'analyse des dialogues produits et l'évaluation de ces systèmes (Jurafsky, 2000). Les dialogues humains comportent beaucoup d'indices qui stimulent l'interaction, cependant nous ne détaillerons ici que ceux qui sont directement utilisés dans les systèmes de dialogue.

Comme décrit dans Jurafsky (2000), les dialogues sont composés d'alternances de prises de parole entre les participants appelées **« tours de parole »**. De ce point de vue, nous pouvons considérer qu'un dialogue est composé d'un ensemble de tours de parole. Les tours de paroles peuvent être composés de séquences très courtes composées d'un brève séquence de mots, à des séquences plus longues qui peuvent durer longtemps. La communication peut même être non verbale, tel qu'un hochement de tête comme fonction phatique, ou tout simplement une expression vocale telle que « hmm ». Dans un dialogue comprenant deux ou plusieurs participants, de nombreux phénomènes incorporés peuvent être observés, par exemple les rapports de force et de hiérarchie, l'initiative etc. Même si le dialogue peut être analysé comme un ensemble de tours de parole, les tours de paroles ne s'enchaînent pas toujours l'un après l'autre mais comportent souvent des chevauchements. Ces indices sont également importants dans la conception des systèmes de dialogue parce qu'ils montrent à un certain degré la frustration du participant lorsqu'il coupe le système (*barge-ins* en anglais) ou lorsque le système lui coupe la parole.

Dans une conversation, les participants peuvent avoir différentes stratégies communicatives. Par exemple, les deux interlocuteurs peuvent avoir une stratégie *coopérative* en tenant compte du but de l'autre ; ou avoir une stratégie *réactive* où le locuteur ajuste son but en fonction de son interlocuteur (Caelen, 2003). Néanmoins, le terme **initiative** est utilisé de façon plus générique pour désigner le locuteur qui guide le dialogue. Dans une conversation humaine, les marqueurs linguistiques et prosodiques ainsi que la communication non verbale permettent cette alternance entre les participants dans les tours de parole (Chu-Carroll et Brown, 1997). Un système de dialogue permettant de donner l'initiative à l'utilisateur au lieu de le guider sur un sujet prédéfini comporte une initiative mixte (Walker et Whittaker, 1990). L'approche formelle de Smith et Gordon (1997) classifie cette notion d'initiative en 4 catégories :

— Directive : Le dialogue est complètement régi par le système, il ne permet pas à l'utilisateur de guider les dialogues ni les sous dialogues

— Suggestive : Le dialogue est guidé par le système, mais permet l'intervention de l'humain sur certaines tâches des sous-dialogues

— Déclarative : L'initiative est à l'utilisateur. Il peut interrompre à tout moment les sous-dialogues

— Passive : L'utilisateur guide le dialogue

Le système de dialogue que nous proposons, comme la plupart des systèmes actuels, doit comporter une initiative-mixte permettant à l'utilisateur d'effectuer une saisie des prescriptions mais aussi d'apporter des modifications en temps réel lorsqu'elles sont nécessaires. Notre proposition s'aligne donc dans la catégorie des systèmes déclaratifs de Smith et Gordon (1997).

2.3 Systèmes de dialogue

Plusieurs termes existent dans la littérature pour désigner une machine capable de dialoguer avec un humain. Le plus populaire est certainement *chatbot* ou *chatterbot*. D'autres termes tels que « agent conversationnel » ou encore « agent conversationnel animé » (*embodied conversational agent*) sont utilisés de façon interchangeable. Pour le sous-ensemble des systèmes de dialogue orientés tâche (ou orientés but), l'utilisation du mot « assistant personnel » est aussi fréquemment utilisée, notamment pour les systèmes commerciaux. Dans ce rapport nous employons le terme « système de dialogue » pour désigner de façon générique tous ces systèmes.

Il est néanmoins important de nuancer ces termes : les chatbots désignent le sous-ensemble des systèmes qui ont pour objectif de dialoguer sur des sujets ouverts. Les défis internationaux comme *Alexa Prize* proposent des compétitions aux universités pour la création de chatbots où sont animées des sessions de conversation d'environ 20 minutes sur des sujets tels que le divertissement, le sport, la technologie, etc.

Le terme agent conversationnel est quant à lui utilisé lorsqu'il y a un agent, un personnage à qui l'on s'adresse. Parfois, ce personnage pourrait être lié à un personnage électronique comme un robot. Les agents conversationnels peuvent être comme des chatbots mais peuvent aussi être orientés tâche comme la plupart des systèmes commerciaux.

Pour ce qui est des agents conversationnels animés, comme leur nom l'indique, ce sont des agents qui proposent un support visuel animé dans l'objectif de créer un attachement plus proche avec l'utilisateur. Dans le domaine de la santé, l'utilisation des agents conversationnels animés est fréquente, notamment dans le cadre des agents développés dans le cadre de la prévention sanitaire auprès de jeunes individus.

Dans cette section, nous allons présenter un état de l'art sur les systèmes de dialogue, en particulier les systèmes de dialogue à base de cadres (*frame-based*) et orientés but. Par la suite nous présenterons un panorama des systèmes de dialogue utilisés dans le domaine de la santé.

2.3.1 Premiers systèmes de dialogue

Les systèmes les plus connus ayant eu un impact historique dans le domaine sont sans doute ELIZA (Weizenbaum, 1966) et PARRY (Colby et coll., 1971). Ces deux chatbots historiques ont une particularité intéressante : ils sont issus du domaine de la psychologie. ELIZA

est conçu pour simuler un entretien avec un thérapeute rogérien qui est fortement basé sur la répétition. L'une des raisons du succès d'ELIZA, hormis son côté précurseur, est lié à l'utilisation d'une technique d'entretien reprenant les énoncés de l'interlocuteur à la place d'avoir recours à une connaissance externe. ELIZA (et PARRY par la suite) est basé sur une approche à base de règles comme d'autres systèmes de leur époque, plus particulièrement une approche de recherche de motifs qui a la capacité de transformer les phrases comme un transducteur à état fini.

Quelques années après ELIZA, le système PARRY, toujours dans le domaine de la psychologie, a été introduit pour étudier la schizophrénie en simulant un patient schizophrène. En plus de la technique de recherche et de remplacement des motifs, Colby a introduit des règles permettant de simuler l'état émotionnel du patient en ajoutant des marqueurs de l'émotion. Le système de PARRY a remporté un succès fabuleux après avoir convaincu un comité de 33 psychiatres humains sur la moitié des dialogues produits. La figure 2.2 montre un exemple de dialogue produit entre un utilisateur (User) et le système. À gauche, le dialogue produit entre l'utilisateur et ELIZA montre bien ce mécanisme de recherche de motif et de remplacement lorsque le système essaie de récupérer des mots clés et répond à l'utilisateur par des syntagmes reformulés par des règles de transformation syntaxique. La partie droite de la Figure 2.2 montre un exemple de dialogue produit entre un utilisateur et le système PERRY. Dans cet extrait traduit à partir de Colby et coll. (1971), les marqueurs de l'émotion sont indiqués entre crochets. Par exemple, le mot clé policier est associé à la mafia dans la graphe et alterne l'état psychologique vers la peur et change le cours du dialogue.

Cette technique de recherche et de remplacement de motifs a donné naissance à beaucoup d'autres systèmes de dialogue en incorporant à chaque fois de nouvelles différences, telle que l'intégration d'un historique personnel (Hutchens, 1996). Cependant, comme l'intitulé de Hutchens (1996) l'indique : «comment passer le test de Turing en trichant», les chercheurs ont commencé à s'intéresser à rentrer dans le sens des énoncés afin de créer des systèmes conversationnels qui ne se limitent pas à la simple répétition d'un mécanisme prédéfini.

En parallèle aux ajouts incrémentaux sur les systèmes à base de règles, les systèmes à base de statistiques ont été introduits dès les années 1980. À la base un générateur de texte probabiliste, le chatbot «Mark V. Shaney », nommé ainsi pour faire référence aux chaînes de Markov (*Markov chain*), est l'un des premiers systèmes à utiliser les techniques probabilistes.

Le deuxième marqueur historique pour les agents conversationnels est peut-être le système «A.L.I.C.E » (Wallace, 2009) introduit par Richard Wallace en 2000. Conceptuellement basé sur une architecture de motif et d'action du système ELIZA, le système ALICE introduit un langage XML appelé *AIML* dédié à la gestion de connaissances des systèmes de dialogue. Jusqu'à présent, beaucoup de systèmes, en particulier commerciaux, sont basés sur cette architecture.

Depuis, les systèmes ont introduit différents traitements issus du TAL comme la tokénisation, la racinisation, l'intégration de Wordnet (Al-Zubaide et Issa, 2011) ou d'une ontologie, etc. De nombreux systèmes hybrides ont également incorporé les motifs avec les modèles de

FIGURE 2.2 – Exemple de dialogue d'ELIZA (Weizenbaum, 1966) et de PARRY (Colby et coll., 1971) traduit en français

Markov (Carpenter, 2007). Les systèmes à base de règles sont utilisés très couramment dans le marché et s'orientent vers l'intégration des technologies du TAL plus sophistiquées dans le même cadre (Tur et De Mori, 2011).

2.3.2 Systèmes de dialogue orientés but

Comme son nom le laisse entendre, un système de dialogue orienté but (aussi dit orienté tâche) est un système qui vise à aider les utilisateurs dans l'achèvement des tâches spécifiques. Le système de dialogue pourrait être conçu pour assister l'utilisateur sur une tâche spécifique comme la livraison de repas à domicile (López-Cózar et coll., 1997) ou sur plusieurs tâches comme la plupart des systèmes de dialogue commerciaux (Ok Google, Siri, Cortana, Alexa, ...) tels que programmer un minuteur, ouvrir une chanson, demander des directions, lire l'actualité, etc. Même s'ils sont orientés pour assister les utilisateurs dans des tâches spécifiques comme celles-ci, certains de ces systèmes sont capables de poursuivre une conversation sociale (chitchat) à petite échelle (Hoy, 2018). Contrairement aux systèmes de dialogue sociaux, ces systèmes se concentrent sur la compréhension efficace de l'utilisateur, la gestion des états du dialogue, et effectuer des actions (liée à la tâche). De même, le nombre de tours de parole devient important et donne une indication significative sur la performance des systèmes (Deriu et coll., 2021), à l'inverse de l'hypothèse qui pense que moins il y a de tours de parole, meilleure serait la performance.

Le domaine d'application des systèmes de dialogue orientés but est vaste, par exemple le

tutorat intelligent (Litman et Silliman, 2004), la réservation de billets d'avion et de train (Lamel et coll., 1999; Hemphill et coll., 1990) ou encore le pilotage d'hélicoptères (Lemon et coll., 2001). Pour assister les utilisateurs dans la réalisation de ces tâches, les systèmes de dialogue ont besoin d'un domaine d'application qui est souvent lié à une ontologie. Le domaine d'application comporte des représentations sémantiques pertinentes pour la tâche ainsi que des intentions qui peuvent être extraites des énoncés des utilisateurs. L'extraction des concepts à partir des cadres sémantiques fait partie d'un processus qui est appelé remplissage d'attributs (*slot-filling* en anglais). Ces notions liées aux systèmes de dialogue à base de cadres seront détaillées dans la section 2.3.3. Historiquement, le système pionnier du domaine est le système GUS (Bobrow et coll., 1977) dans le domaine de réservation.

L'exemple de dialogue présenté dans la Figure 2.3 de Bobrow et coll. (1977) montre un dialogue produit dans le système de dialogue orienté-but à initiative mixte dans le cadre de réservation des billets d'avion.

(1) GUS: Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go?
(2) CLIENT: I want to go to San Diego on May 28
(3) GUS: What time do you want to leave?
(4) CLIENT: I must be in San Diego before 10 am
(5) GUS: Would you like P.S.A flight 102 that arrives at 9.15 am?
(6) CLIENT: What is the next flight?
(7) GUS: Air California flight 310 that leaves at 8.30 am and arrives at 10.00 am
(8) CLIENT: I'll take the first one
(9) GUS: In what name should I make a reservation, please?
(10) CLIENT: Dan Bobrow
(11) GUS: I have confirmed the following flight : P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.
(12) GUS: What date do you want to return on?
(13) CLIENT: On Friday in the evening.
(14) GUS: Would you like the flight that leaves at 7.45 pm?
(15) CLIENT: That's fine.
(16) GUS: I have confirmed the following flight : P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm. Thank you for calling. Goodbye.

FIGURE 2.3 – Extrait d'un dialogue orienté but dans le domaine de réservation de billets d'avion (Bobrow et coll., 1977)

Dans cet exemple, on voit que le dialogue démarre avec l'initiative du système **(1)** qui invite l'utilisateur à formuler une demande liée à la réservation de billets d'avion. Dans les tours de parole**(2-4)**, l'utilisateur formule sa demande, et le système extrait les concepts et les valeurs lié au domaine d'application (ville-départ, ville-arrivée, heure, ...) Il est à noter qu'un système orienté but est amené assez souvent à faire des requêtes à une base de connaissances afin de poursuivre le dialogue, par exemple ici pour récupérer les vols correspondants aux critères de l'utilisateur. Dans **(6)** l'utilisateur initie un sous-dialogue en demandant les prochains vols sans changer de « but » principal. Dans **(11 et 16)** le système répète les valeurs capturées par les cadres afin de vérifier implicitement les informations et d'éviter qu'il y ait

des erreurs.

2.3.3 Systèmes de dialogue à base de cadres

Les systèmes à base de cadres aussi appelés à base de schémas (*form based*) représentent les données dans une structure de cadres. Ces systèmes proviennent de la théorie de la sémantique de cadres introduite par Fillmore et Baker (2001). Cette structure permet de récupérer au fur et à mesure des bouts d'informations de l'utilisateur dans un système de dialogue transactionnel. En effet, dans un système orienté tâche, le but est de réaliser une action lorsque les informations nécessaires sont récupérées de l'utilisateur. Cela nécessite de définir un champ sémantique bien défini et restreint à un domaine spécifique.

Inspiré des théories du langage décrites dans 2.2.1, ces systèmes traitent le problème en deux parties : d'une part le système remplit des attributs (*slots* aussi appelés *frame elements*) qui correspondent à des bouts d'informations pertinentes à la définition du domaine et aux champs sémantiques et il en extrait les valeurs. D'autre part, étant donné qu'il s'agit d'un énoncé, le système identifie le cadre sémantique, c'est-à-dire l'intention de l'énoncé.

L'intention de l'énoncé est une notion liée au domaine d'application d'un système de dialogue orienté tâche. De manière générale, ces systèmes sont conçus pour assister les utilisateurs à réussir une tâche dans un domaine bien défini : par exemple, la réservation d'un billet d'avion. Lorsqu'un système comporte des fonctionnalités concernant différents domaines, les intentions de l'énoncé de chaque domaine peuvent être différentes ou identiques les unes des autres, mais elles ont toutes pour but de réussir les tâches de chaque domaine.

L'utilisation des cadres a d'abord été utilisée dans le domaine de la compréhension automatique du langage où le sens de l'énoncé est obtenu en saisissant des attributs relatifs à la sémantique de l'énoncé. L'exemple de dialogue produit dans le domaine de réservation de billets d'avion était présenté sur la Figure 2.3. Dans cet exemple, lorsque l'utilisateur demande s'il y a des jours d'une destination à une autre, le système l'associe à l'intention *SHOW-FLIGHTS*. La figure 2.4 schématise le processus de remplissage d'attributs dans le cadre de réservation de billets.

$$
\begin{aligned}
\mathit{SHOW\_FLIGHTS}(\ & \text{ORIGIN\_CITY} = \textsc{Boston},\\
& \text{ORIGIN\_DATE} = \textsc{mardi},\\
& \text{ORIGIN\_TIME} = \textsc{matin},\\
& \text{DEST\_CITY} = \textsc{San Fransisco})
\end{aligned}
$$

FIGURE 2.4 – Représentation des concepts dans le domaine des réservations de billets d'avion

Contrairement à un système à état fini, un système à base de cadres permet de communiquer des informations dans un ordre et des combinaisons différentes. L'une des raisons pour lesquelles ces systèmes sont appelés « à base de schémas » est liée à cette caractéristique permettant au système de compléter des informations pour remplir les attributs nécessaires à

partir des templates prédéfinis. Ainsi, lorsque l'utilisateur tente de surinformer le système en précisant des informations déjà connues par le système, le système n'a pas besoin de créer des états spécifiques.

L'autre avantage est la possibilité de créer plus facilement des systèmes qui ont une initiative mixte. L'utilisation des cadres pour représenter des données facilite l'attribution de l'initiative à l'utilisateur. Cependant, contrairement aux dialogues humains, l'interaction avec un système de dialogue est souvent composée de tours de parole. Dans un système orienté tâche, chaque tour de parole est composé d'un énoncé et d'une intention, que ce soit ceux de l'utilisateur ou du système.

2.3.4 Systèmes de dialogue de santé : un domaine émergent

Dans le domaine de la santé, la plupart des systèmes se focalisent sur la prévention de santé des patients, notamment dans le cadre de la santé mentale (Fitzpatrick et coll., 2017; Miner et coll., 2016; Hudlicka, 2013) mais d'autres domaines d'applications existent tels que le diagnostic préventif (Beveridge et Fox, 2006; Philip et coll., 2014), le suivi ambulatoire (Giorgino et coll., 2005), la collection de données de santé (Black et coll., 2005), etc. Les revues de littératures et les experts du domaine soulignent le manque de méthode d'évaluation qui permettrait de voir l'efficacité et la sécurité de ces systèmes, en particulier sur le long terme (Laranjo et coll., 2018; Kocaballi et coll., 2019). Cependant, malgré un manque de méthode d'évaluation standardisée, des essais aléatoires et contrôlés montrent l'efficacité et l'acceptabilité de l'utilisation des conversationnels (Gaffney et coll., 2019; Bibault et coll., 2019; Fitzpatrick et coll., 2017). Les systèmes de dialogue que nous avons cités se concentrent sur la réalisation du dialogue avec les patients, cependant il existe des systèmes qui sont conçus pour les experts médicaux. Par exemple, Llanos et coll. (2015); Campillos-Llanos et coll. (2020) présente un système de dialogue qui simule un patient lors d'une consultation médicale dans un but pédagogique. Dans le processus thérapeutique, une relation solide entre le patient et le thérapeute se fonde sur un principe de confiance partagée qui se développe de manière générale au fil des séances. Cette relation exige de l'empathie et de la compréhension pour parvenir à une décision commune, en traitant souvent de grandes zones d'incertitude et en équilibrant des risques concurrents (Bell et coll., 2019; Powell, 2019). L'un des verrous scientifiques pour les systèmes de dialogue en santé, ou de manière générale de l'utilisation de l'intelligence artificielle dans le domaine médical est le manque de cette empathie, de cette confiance et du sens des relations entre le système conversationnel et le patient (Bell et coll., 2019; Powell, 2019). Malgré la complexité du processus de la création d'une relation solide fondée sur une confiance partagée, des travaux récents abordent le sujet de ce point de vue (Morris et coll., 2018).

En effet, utilisés dans le domaine de la psychiatrie, les systèmes de dialogue montrent une efficacité prometteuse (Vaidyam et coll., 2019; Bibault et coll., 2019). Par exemple, le travail de Bibault et coll. (2019) compare un système de dialogue, le système «Vik» [3], et des clini-

3. https ://www.wefight.co/fr/accueil/

ciens pour l'annonce et l'assistance à des patientes atteintes du cancer de sein. Les résultats montrent que l'assistance donnée aux patientes par le système de dialogue n'était pas inférieure à celle donnée par les experts médicaux sur des sujets simples. En effet, les systèmes de dialogue peuvent être bénéfiques, notamment dans le cadre de la prévention de santé mineure. Les auteurs supposent qu'il serait même envisageable de prescrire «un chatbot» afin d'assurer le suivi ambulatoire. Cette idée rejoint la conclusion d'autres systèmes de dialogue développée dans le cadre du suivi de la santé mentale.

2.4 Architecture d'un système de dialogue oral à base de cadres

Les principes décrits sur les systèmes de dialogue, notamment ceux orientés but sont génériques pour des conversations textuelles et orales. D'autres types de systèmes de dialogue existent, par exemple les systèmes multimodaux qui font partie des défis actuels de la gestion de l'état du dialogue 9(DSTC) (Gunasekara et coll., 2020) ou systèmes de dialogue incarnés dans la réalité virtuelle (Brinkman et coll., 2012). Malgré les petites différences selon les modalités d'entrée et de sortie (oral, virtuel, textuel, embodied), l'architecture globale d'un système de dialogue actuel reste similaire. En revanche, nous pouvons distinguer trois grandes catégories de systèmes de dialogue :

— Systèmes à base de règles

— Systèmes modulaires (pipeline)

— Systèmes de bout en bout

Les systèmes à base de règles sont les plus historiques basés sur les automates à états finis et déterministes. Les systèmes présentés dans la section 2.3.1 font partie de cette grande catégorie des systèmes qui sont toujours utilisés dans certains systèmes commerciaux. Les systèmes modulaires et les systèmes de bout en bout utilisent des méthodes à base d'inférence, appris sur un ensemble de données. L'architecture d'un système de dialogue modulaire typique est schématisée sur la Figure 2.5. Dans une approche modulaire, le traitement suit un schéma de traitement (pipeline) qui commence par la reconnaissance de la parole de l'utilisateur jusqu'à la production de la voix artificielle par le système.

Cette section présentera ces 3 différentes grandes approches à la conception des systèmes de dialogue dans la partie 2.4.1. Dans un système de dialogue oral, le schéma de traitement (pipeline) démarre avec la reconnaissance automatique de la parole (RAP) de l'utilisateur. Le module de RAP transforme le signal de la parole d'un énoncé (*utterance* en anglais) en une transcription textuelle de l'énonciation. Sujette à erreurs, le module de compréhension du langage (NLU) extrait les valeurs des cadres à partir de cette transcription ainsi que l'intention du locuteur. Deuxième étape de la Figure 2.5 montre pour le domaine de réservation de billets, les concepts extraits (slots) sous forme d'étiquette-valeur et valeurs. Cette représentation de sens et l'intention extraite à partir de l'énoncé transcrit sont celles qui sont

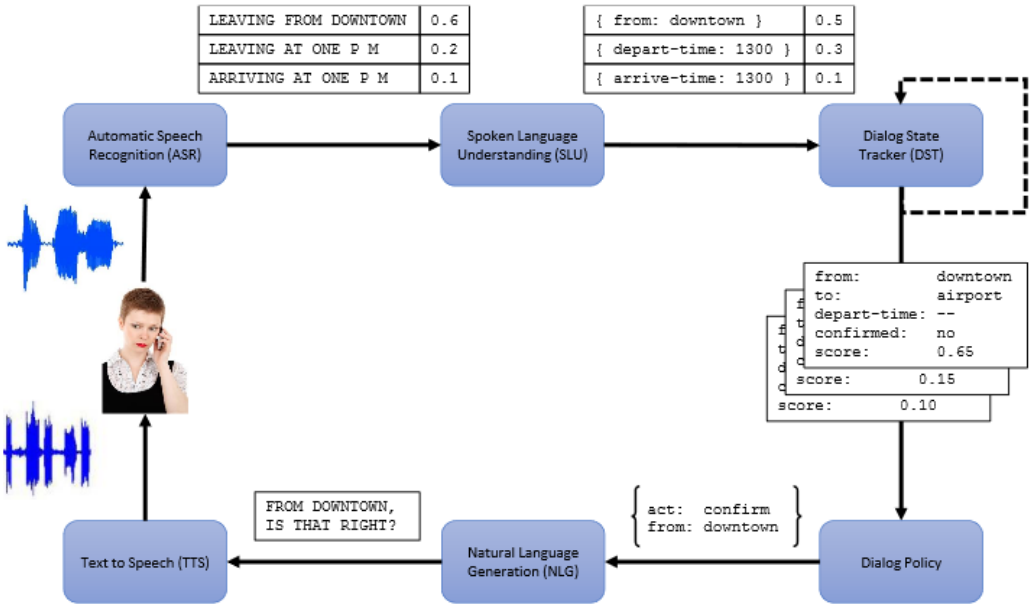


FIGURE 2.5 – Architecture classique d'un système de dialogue orienté but modulaire (Williams et coll., 2016)

passées au module de suivi de l'état du dialogue (DST). Ce module est aussi appelé comme le gestionnaire de dialogue et permet au système de déterminer l'état du dialogue. Cette tâche consiste à inférer correctement l'état de la conversation, tel que l'objectif de l'utilisateur, compte tenu de l'historique du dialogue jusqu'au tour de dialogue actuel (Williams et coll., 2016). Étant donné l'état de la conversation, le système doit répondre à l'utilisateur. Le choix de la réponse qui va être donné à l'utilisateur est déterminé par la politique de dialogue qui consiste à choisir l'action du système adéquat à l'état du dialogue. L'action déterminée par la politique de dialogue est souvent intrinsèque au dialogue permettant de déterminer la réponse qui sera donné à l'utilisateur. Souvent, dans les systèmes de dialogue orientés tâches, la réalisation de la tâche nécessite la communication avec une base de connaissances ou un API pour pouvoir correctement répondre à l'utilisateur. L'action peut-être donc un appel à une base de connaissances, par exemple pour récupérer les vols disponibles dans le contexte du domaine de réservation de billets d'avion. Dans un cadre multimodal, l'action pourrait consister à afficher visuellement une information. Dans cette thèse, nous distinguons les deux comme les actions internes et externes au dialogue. Par la suite, l'action du système est transformée en langue naturelle par le module de la génération automatique du texte (GAT). En fonction de l'action déterminée par la politique et les concepts associés, le module de GAT produit de langue naturelle à partir de la représentation de sens sortant de la politique de dialogue. Enfin, la dernière étape du schéma consiste à transformer l'énoncé produit par le module de GAT en un signal de parole par le module de la synthèse vocale. Pour un système multimodal, cela pourrait être un affichage visuel à la place d'une voix produite. Pour tous ces modules, il existe une grande variété d'approches, notamment stochastiques, que nous allons brièvement présenter dans les sous-sections 2.4.2 à 2.4.7.

2.4.1 Familles d'architecture des systèmes de dialogue

Les premiers systèmes de dialogue que nous avons abordés dans la section 2.3.1 s'appuient sur des méthodes à base de règles basées sur les transducteurs à états finis et déterministes. La Figure 2.6 illustre avec un exemple simple le principe des systèmes à états finis. La portion du dialogue illustrée sur cet exemple démarre avec l'initiative du système. En effet, un système à base de règles cherche à récupérer les informations au fur et à mesure selon la définition des états. La formalisation d'un automate à état fini pourrait être résumée comme ce qui suit :

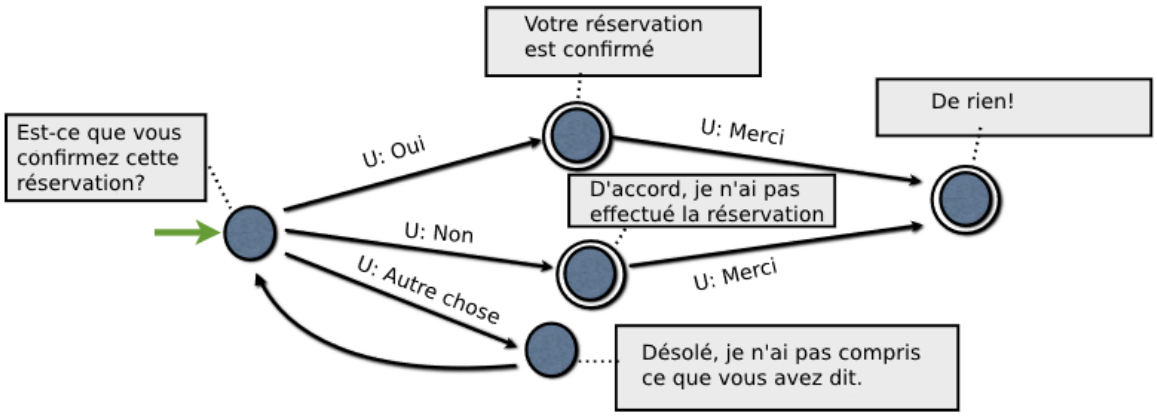


FIGURE 2.6 – Exemple du fonctionnement d'un système de dialogue basé sur automate à état fini

— Ensemble fini et non vide d'états S (atomiques), dont chaque état est associé à une action spécifique du système

— Un ensemble fini et non vide Σ d'entrées d'utilisateur possibles

— Une fonction (partielle) $\delta : S\ x\ \Sigma \longrightarrow S$ définissant la transition entre chaque état accepté par l'automate

— Un état initial $s_0 \in S$

— Un ensemble d'états finaux $F \subset S$

La formalisation ci-dessus définit le principe de fonctionnement d'un automate à état fini modélisant les dialogues dans le cadre de la réalisation d'une tâche spécifique. Les transitions peuvent correspondre à des tâches extra-linguistiques telles que l'interrogation d'une base de connaissance. Comme dans ELIZA et PARRY, ces systèmes peuvent comporter des règles de transformation de chaîne de caractères, implémenter un mécanisme de mémoire et de seuil de confiance, etc. afin d'avoir des dialogues efficaces en fonction du contexte. Dans un système de dialogue oral et déterministe, le graphe d'états peut être converti en un script « VoiceXML » (Lucas, 2000). Historiquement introduit par AT&T et plus tard développé par IBM et Motorola, ce langage de balisage permet la mise en place d'une application vocale sur internet intégrable aux systèmes de RAP et de TTS existants. Dans ces systèmes les transitions vers les états suivants sont prédéterminées et sont basées sur de des simples

conditions, telles que le degré de confiance. L'utilisation de graphes convient aux interactions assez simples où le déroulement du dialogue peut être prédéterminé. L'avantage principal de ces systèmes repose sur la facilité de mise en place. Vu qu'il s'agit d'un ensemble de règles, ce sont des systèmes relativement rapides à mettre en place et qui produisent des dialogues prédictibles (pour le concepteur dialogique ainsi que l'utilisateur final). Cet aspect est industriellement important, car cela rend la gestion d'erreurs plus facile notamment pour comprendre la provenance d'une erreur. Néanmoins, ces systèmes nécessitent une analyse très fine de l'interaction conversationnelle au niveau de la création de règles. Face aux incertitudes, notamment dans le cadre d'un système de dialogue oral où il peut y avoir des erreurs de reconnaissance de la parole, les règles déterministes ne sont pas adaptées. Face à ces défis-là, la deuxième grande catégorie des systèmes de dialogue a émergé : les systèmes de dialogue à base de cadres, introduits dans la section 2.3.3.

Ces systèmes peuvent être vus comme une généralisation de l'approche à base d'états finis permettant à l'utilisateur de « remplir » les informations par de multiples chemins. Dans ces systèmes l'utilisateur peut répondre aux questions du système, mais il peut aussi donner d'autres informations que le système n'avait pas demandées. Cette caractéristique est liée à la conception des cadres sémantiques (Fillmore et coll., 1976). Dans cette thèse, nous nous sommes basés sur une approche à base de cadres. En revanche, lorsque la tâche à réaliser est complexe, il est possible qu'un cadre ne soit pas suffisant pour modéliser la tâche. Par exemple, si nous prenons l'exemple de la réservation de billets d'avion et l'appliquons à la réservation de billets d'autres transports, les informations pertinentes seraient différentes. Pour la modélisation des tâches complexes comportant plusieurs cadres, une approche consiste à effectuer la gestion du dialogue basée sur un agenda (Rudnicky et Xu, 1999; Xu et Rudnicky, 2000). La Figure 2.7 extrait de Rudnicky et Xu (1999) illustre le schéma de fonctionnement de la gestion de dialogue basée sur un agenda.

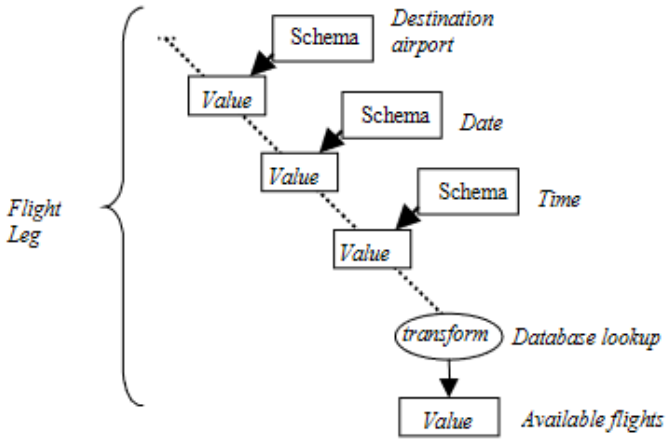


FIGURE 2.7 – Description de la tâche de la réservation de billets d'avion en une succession de sous-tâches en utilisant un agenda (Rudnicky et Xu, 1999)

Schématisé sur la Figure 2.7, le système basé sur l'agenda est représenté sous la forme d'un arbre, qui reflète la hiérarchie naturelle et l'ordre des informations nécessaires à la réa-

lisation de la tâche (Rudnicky et Xu, 1999). Au cours d'une session, l'utilisateur peut ajouter des étapes à un voyage plutôt que de travailler à partir d'un schéma fixe. Le fonctionnement de la gestion basée sur un agenda repose sur l'utilisation d'une structure de données arborescente modifiable dynamiquement au fur et à mesure des objectifs de l'utilisateur et du système. La création de règles étant un processus chronophage et difficile lorsque la tâche est complexe, la communauté s'est tournée vers les modèles quantitatifs permettant l'apprentissage de stratégies dialogiques ainsi que la gestion du dialogue. Les premiers travaux stochastiques étaient basés sur les processus de décision markoviens (MDP) (Young, 2000; Levin et coll., 2000). L'élaboration de règles est basée principalement sur l'heuristique ainsi que les essais et erreurs (Levin et coll., 2000). Il est généralement difficile de prédire toutes les situations conversationnelles possibles que le système pourrait rencontrer dans des scénarios réels. De plus, lorsqu'une nouvelle application est développée, il faut souvent recommencer tout le processus de conception depuis le début. Pour les MDP, la conception du système est vue comme un problème d'optimisation et est modélisée comme un processus de décision séquentielle composé d'un ensemble d'actions, d'états et de stratégies. En revanche, dans un système de dialogue, les entrées utilisateurs viennent d'un signal de parole où les erreurs de reconnaissance vocale sont courantes (Williams et coll., 2008). Cela rend difficile le choix de l'action du système étant donné que l'état peut-être erroné. De même, au cours du dialogue, les intentions utilisateurs peuvent changer à tout moment. En conséquence, la machine doit décider si les preuves contradictoires ont été introduites par une erreur de reconnaissance de la parole, ou par une nouvelle intention de l'utilisateur. Pour toutes ces raisons, Young et coll. (2010) propose que la gestion du dialogue soit considérée comme une planification dans l'incertitude par les processus de décision markoviens partiellement observables (POMDP). Contrairement aux processus markoviens classiques, dans ces systèmes l'agent a une croyance (*belief* en anglais) au lieu d'une connaissance déterministe de l'état courant. Cette notion de croyance est introduite pour gérer la notion de l'incertitude face aux potentielles erreurs et montre qu'il permet d'obtenir de meilleures performances dans des cas d'usages réels (Young et coll., 2016).

Selon Zhao et Eskenazi (2016), l'approche modulaire à un système de dialogue comporte deux inconvénients majeurs. En premier, il s'agit de la satisfaction utilisateur et des constatations d'erreurs qui portent sur la globalité du système lorsqu'un interlocuteur interagit avec le système. Afin d'améliorer le système et corriger des erreurs, il est nécessaire d'analyser finement ces différents modules afin d'effectuer des changements. Deuxièmement, il est difficile de changer un composant indépendamment des autres. Par exemple, l'introduction des nouvelles données au module de la compréhension du langage nécessite le réentraînement de la politique du dialogue qui prend en compte ces nouveaux types de données (Zhao et Eskenazi, 2016). En conséquence, la communauté s'est tournée très rapidement vers des méthodes de bout en bout (*end-to-end*). Ce changement de tendance est visible à partir du 6ème défi de DSTC (Perez et coll., 2017) qui montre l'intérêt de la communauté sur les systèmes entraînables de bout en bout. Le fonctionnement de ces systèmes repose sur l'apprentissage supervisé sur des grandes quantités de données dialogiques. Zhao et Eskenazi

(2016) a montré qu'il est possible d'entraîner conjointement la gestion de suivi (DST) et la politique du dialogue. L'objectif d'un système bout en bout est d'approximer une fonction capable d'associer le signal d'entrée (énoncé) à un énoncé de sortie en apprenant sur une grande quantité de données.

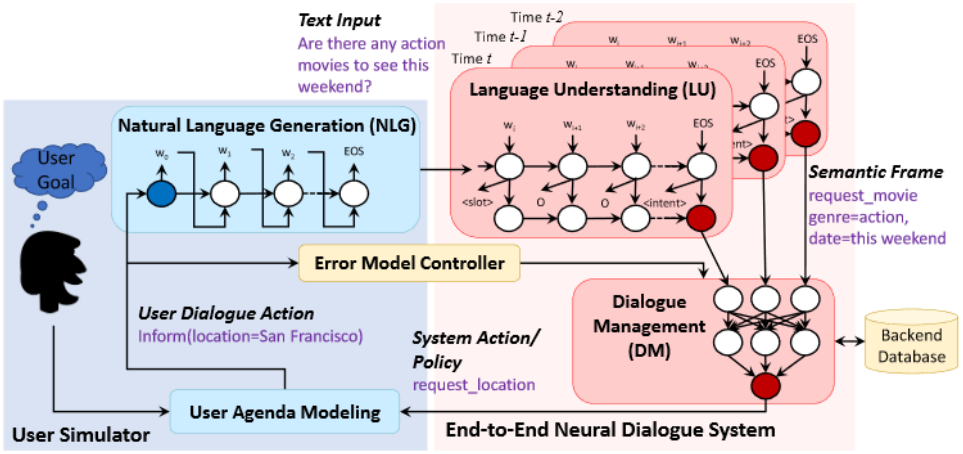


FIGURE 2.8 – Gestion de dialogue de bout-en-bout (Li et coll., 2017)

La Figure 2.8 schématise ce processus de gestion de dialogue de bout en bout. Contrairement à une approche modulaire, sur la Figure 2.8, l'apprentissage est fait en un module composé de plusieurs couches dans un cadre d'apprentissage par renforcement. La compréhension du langage est effectuée dans une couche cachée et la représentation intermédiaire est passée au gestionnaire de dialogue qui détermine l'état de suivi du dialogue et la politique. Il s'en suit que la sémantique du domaine est représentée par un espace latent et n'est pas explicite. Les modèles de bout en bout sont entraînés sur des corpus de dialogue et sont capables d'associer les énoncés à une réponse du système ou une action externe lorsque le système a recours à une base de connaissance. Cette association directe des énoncés aux actions comporte un avantage : le modèle apprend les représentations intermédiaires de la sémantique dans un espace latent qui est optimisé pour la tâche.

Si les modèles de bout en bout sont particulièrement performants pour des tâches ayant suffisamment de corpus d'apprentissage, ils restent toutefois difficiles à transférer à une application industrielle nouvelle. En effet, dans l'industrie l'objectif est de construire des modèles agiles, prévisibles, durables et évolutifs. Par conséquent, les approches statistiques doivent permettre un contrôle explicite sur le *pipeline* de traitement. En particulier dans le domaine de la santé, sans utiliser de mécanisme de contrôle intermédiaire, déduire directement une sortie à partir d'une représentation d'entrée peut avoir de graves conséquences (Kocabiyikoglu et coll., 2020). Le développement d'un système de dialogue dans le domaine de la santé implique des exigences spécifiques, notamment pour répondre aux normes de sécurité, d'efficacité et de traçabilité des décisions des logiciels. Par ailleurs, un système industriel exige de pouvoir être modifié rapidement suite à des erreurs logicielles ou de nouvelles régulations.

Par exemple, En France, L'Agence Nationale de Sécurité du Médicament (ANSM) met à jour une liste de médicaments autorisés sur le territoire. Un médicament peut perdre son

autorisation de mise sur le marché (AMM) suite à une décision européenne ou nationale. Dans une approche de bout en bout, il est difficile d'introduire un mécanisme de contrôle qui pourrait incorporer dynamiquement ce type de modifications dans un LAP. De plus, un système doit être entièrement traçable. En cas d'erreur, il faut être capable d'identifier facilement la source de cette dernière. Le découplage des tâches dans un système modulaire le permet plus facilement que dans un système de bout en bout en créant des *logs* de chaque module.

Enfin, dans l'industrie, il n'est pas rare d'« intervenir » lorsqu'il y a un problème avec un logiciel. Par exemple, en modifiant manuellement le comportement d'un logiciel ou en corrigeant une entrée dans la base de données lorsqu'un client a un besoin spécifique. Dans ce contexte, les réseaux de neurones sont souvent complexes et difficiles à prévoir, plus difficiles à tester, à expliquer et à maintenir (Sculley et coll., 2015). Ils sont considérés comme des boîtes noires dans le sens où, bien qu'ils puissent estimer n'importe quelle fonction, leur structure ne fournit pas d'indications sur la fonction estimée. Ainsi, les solutions rapides et même l'ajustement léger du système en fonction de nouvelles données nécessitent le réentraînement de l'ensemble du système.

Pour les dialogues, la première adaptation d'un système de bout en bout était réalisée à partir de l'architecture encodeur-décodeur (Sutskever et coll., 2014) sur les systèmes de dialogue sociaux avec succès (Vinyals et Le, 2015; Shang et coll., 2015). Cette famille d'approches traite le problème de dialogue comme un problème de transduction de séquences à partir des énoncés (Wen et coll., 2016). Les énoncés des utilisateurs sont encodés dans des vecteurs représentant implicitement la sémantique qui sont associés par le décodeur à des réponses du système. Ils sont majoritairement utilisés dans les systèmes de dialogue sociaux car il est difficile d'injecter une connaissance spécifique liée à un domaine comme l'interaction avec une base de connaissance (Yin et coll., 2015; Wen et coll., 2016). Pour adapter ces systèmes aux dialogues orientés tâche, il y a eu plusieurs propositions. Le premier volet de ces systèmes se base sur les réseaux de mémoires (*memory networks* en anglais) (Weston et coll., 2014; Sukhbaatar et coll., 2015). Cette approche adaptée pour les systèmes de dialogue orientés tâche par Bordes et coll. (2016) sauvegarde un historique dialogique permettant au système d'avoir recours à cette connaissance sous forme de mémoire à court terme qui peut être utilisée pour le raisonnement et la génération de réponses. Afin de prendre en compte les spécificités du dialogue orienté tâches, Bordes et coll. (2016) inclut dans le vocabulaire des mots clés représentant un appel à la base de connaissance avec les mots retenus dans le réseau de mémoire. D'autres systèmes fondés sur le principe de réseaux de mémoire ont été proposés : les réseaux de mémoire à porte (gated) (Liu et Perez, 2017), séquence vers séquence (seq2seq), réseaux avec un mécanisme de copie (copy-augmented) (Eric et Manning, 2017), etc. Cependant, tous ces systèmes nécessitent des corpus de dialogue assez grands pour l'entraînement. Les informations de la base de connaissance sont extraites à partir des corpus de dialogues existants. Les appels aux ressources externes sont limités par un modèle entraîné hors ligne, et qui peut être difficilement applicable aux interactions des utilisateurs réels (Liu et Lane, 2018). Face à ce problème, différentes propositions existent :

Wen et coll. (2016) propose un réseau de neurones entraînable de bout en bout mais de façon modulaire qui demande une connaissance explicite des concepts clés et valeurs liés à une base de connaissance. D'une façon similaire, Williams et coll. (2017) propose une architecture permettant l'incorporation de la connaissance procédurale via des templates d'actions proposés dans l'entraînement. Il s'agit d'une approche à base de réseaux de neurones récurrents (RNN) avec un modèle à base de règles entraînable en deux parties, d'abord un apprentissage supervisé, puis par un apprentissage par renforcement. Cependant, la plupart de ces applications traitent les réseaux de neurones comme une « boîte noire » vu qu'il est très difficile de comprendre et d'interpréter les représentations cachées (Annasamy et Sycara, 2019).

En effet, l'une des critiques faites sur les systèmes de dialogue composés de plusieurs sous-composants est la difficulté d'entraîner et d'optimiser les modules indépendamment les uns les autres (Serban et coll., 2016; Liu et Lane, 2017). Comme les résultats des différents systèmes font partie de l'entrée d'autres composants, les erreurs faites ultérieurement sont propagées dans les autres modules. Par conséquent, dans des systèmes composés de plusieurs composants, le NLU occupe une place primordiale dans le bon fonctionnement du système.

2.4.2 Reconnaissance automatique de la parole

Le premier composant d'une chaîne de traitement classique de dialogue commence par la voix. Dans un système oral, il est nécessaire de reconnaître la voix de l'utilisateur. Cette première étape permettant de transformer le signal de parole en une transcription est gérée par un module de la reconnaissance automatique de la parole (RAP). La sortie de la RAP est en général un énoncé associé à un score de confiance ou une liste de n-meilleurs énoncés. Indépendamment de la sémantique de cadres, ces scores de confiance permettent la gestion de l'incertitude à un certain degré par le biais des actes de paroles déterminées selon les différents niveaux de confiance.

Les approches récentes de la reconnaissance automatique de la parole s'appuient sur des méthodes stochastiques apprises sur des corpus alignés (parole-transcription). Comme pour les systèmes de dialogue sociaux (Vinyals et Le, 2015; Shang et coll., 2015), l'approche classique neuronale s'appuie sur les méthodes de séquences à séquences (seq2seq) (Chorowski et coll., 2014; Chan et coll., 2016; Bérard et coll., 2016). Similaire à l'abstraction de la sémantique pour les systèmes de dialogue de bout en bout, l'utilisation d'une architecture encodeur-décodeur avec de l'attention permet pour un système de RAP de s'en passer d'une décomposition en phonèmes. De cette manière, un signal sonore peut-être directement associé à une transcription comparée aux approches plus classiques comme Povey et coll. (2011). Ces systèmes de RAP conventionnels sont composés d'un modèle acoustique, d'un modèle de prononciation et d'un modèle de langue qui sont entraînés séparément. Dans ces systèmes, la constitution d'un lexique de prononciation et la définition de l'ensemble de phonèmes pour une langue particulière nécessitent des connaissances spécialisées et sont

chronophages.

2.4.3 Compréhension automatique du langage naturel

Le deuxième composant principal est celui de la compréhension automatique du langage naturel (NLU). Lorsque le système est oral, il s'agit de la compréhension de la langue orale (SLU) qui comporte des phénomènes de la parole spontanée. Historiquement, les premiers systèmes de SLU étaient basés sur la compréhension des transcriptions de NLU. Toutes deux consistent à transformer des énoncés en une représentation formelle. Les systèmes de SLU récents se concentrent sur l'association de la parole directe à l'intention ainsi qu'aux cadres sémantiques de façon de bout en bout (Lugosch et coll., 2019; Desot et coll., 2019a). La NLU permet ainsi de créer une représentation intermédiaire comportant l'intention de l'utilisateur et les cadres qui sont identifiés dans l'énoncé. Comme dans le module de RAP, l'intention ainsi que les attributs peuvent être associés à des scores de confiance. Ce composant étant la partie essentielle permettant l'extraction de la sémantique des énoncés, nous le détaillerons dans la section 2.5.

2.4.4 Gestion de l'état du dialogue

Toutes les démarches jusqu'à maintenant permettent d'extraire les cadres ainsi que l'intention de l'utilisateur dans un contexte bien défini, indépendamment d'un contexte dialogique. Dans la section 2.2, nous avons abordé les caractéristiques des conversations humaines et avons mis l'accent sur les phénomènes complexes de la nature de la conversation humaine. Ce composant de gestion de l'état du dialogue (*Dialogue State Tracking (DST)*) de l'architecture permet d'ajouter cette abstraction en associant un état au dialogue en cours. Il effectue la suivi de l'état actuel des cadres ainsi que l'acte de dialogue le plus récent de l'utilisateur.

Grâce à ce module, l'état du dialogue dans l'instant t ne comporte donc pas seulement la représentation sémantique de l'énoncé, mais inclut également l'état complet des cadres en faisant une sorte de résumé du contexte dialogique. Ce composant est aussi chargé de récupérer toutes les informations nécessaires dans un système orienté tâche. Par rapport aux attributs saisis par l'utilisateur, des questions prédéfinies sont orientées vers l'utilisateur afin de récupérer tous les attributs nécessaires pour effectuer la tâche.

Différentes approches neuronales sont proposées dans le domaine du DST telles que les approches à base de RNN (Henderson et coll., 2014; Wen et coll., 2016), approche à base d'attention (Wu et coll., 2019; Xu et Hu, 2018) et plus récemment des méthodes à base de l'architecture transformeur (Zhang et coll., 2019a; Heck et coll., 2020; Rastogi et coll., 2020).

2.4.5 Politique du dialogue

Lorsque nous interagissons avec quelqu'un, un moyen simple de savoir si le message transmis est compris par l'entendeur est sa réaction face au message. Essentiellement, pour

assurer l'intercompréhension, nous utilisons des stratégies pour trouver un terrain commun. Dans une architecture modulaire, cette particularité du dialogue est représentée par une action prise par rapport à l'état du dialogue.

De manière formelle, celle-ci peut être représentée comme dans l'équation 2.2 :

$$\hat{A}_i = \underset{A_i \in A}{\operatorname{argmax}} P(A_i | A_1, U_1, ..., A_{i-1}, U_{i-1}) \tag{2.2}$$

Étant donné le tour i, l'objectif est de prédire l'action A_i en tenant compte de l'ensemble des états du dialogue (Jurafsky, 2000). L'état étant composé des actes de parole du système A et les actes de parole de l'utilisateur U, la tâche est de trouver l'action la plus probable.

Comme les autres composants du système de dialogue, la politique est aussi associée à la théorie des cadres. Lorsqu'il y a une action qui est prédite à partir de l'ensemble des états du dialogue, les attributs saisis sont véhiculés avec l'action prédite. Si nous prenons pour exemple un système de réservation de billets, si l'action prédite est d'effectuer la réservation pour telle date, la valeur de la date saisie dans l'attribut serait transmise dans l'action. La politique du dialogue peut-être entraînée de façon supervisée ou par par renforcement (Chen et coll., 2018). L'approche classique consiste à effectuer une correspondance entre le contexte du dialogue aux actions du système par des règles créées manuellement. La création de règles peut-être également réalisée par un agent permettant de démarrer le système puis apprises automatiquement de façon supervisée (Chen et coll., 2017). Les règles sont spécifiques à la tâche, difficiles à créer lorsque la tâche est complexe et difficile à mettre à jour dans le long terme. Dans un cadre supervisé, la politique du dialogue peut-être vue comme une tâche de classification classique et entraînée de façon supervisée avec les réseaux de neurones (Henderson et coll., 2005, 2008; Williams et Zweig, 2016).

En revanche, l'apprentissage supervisé comporte aussi des désavantages. L'inconvénient principal de l'approche supervisée est lié à la représentation des données de l'échantillon dont on dispose. Les dialogues collectés peuvent être réduits à certains contextes et parfois ne pas couvrir toutes les situations dialogiques possibles. Ce phénomène de représentation de données n'est pas spécifique à l'apprentissage de la politique du dialogue, mais s'applique sur tous les composants stochastiques et l'apprentissage de bout en bout. Ce phénomène bien connu dans le domaine du dialogue est dénommé comme le problème classique de la paradoxe de l'œuf et de la poule. L'une des premières propositions face à ce problème est celle de Schatzmann et coll. (2007) qui consiste à générer des données en simulant un utilisateur basé sur un agenda pour démarrer (*bootstrap* en anglais) un système POMDP. Pour l'apprentissage de la politique du dialogue, la politique apprise par renforcement pourrait être considérée dans la même optique. Avec l'apprentissage par renforcement, il est possible d'entraîner une politique sans avoir besoin d'initialisation experte qui est optimisée automatiquement à partir des interactions réelles (Gašić et Young, 2013; Su et coll., 2017). Dans les approches par renforcement, chaque tour de dialogue reçoit une mesure de performance appelée comme la récompense (*reward* en anglais) et l'agent explore les différentes séquences d'actions dans différentes situations et effectue des ajustements afin de maximi-

ser la somme attendue des récompenses (Williams et coll., 2017). Différentes approches par renforcement sont proposées dans la communauté dans le domaine de l'entraînement de la politique du dialogue (Peng et coll., 2018; Dhingra et coll., 2016; Williams et coll., 2017). Cependant, dans ces méthodes, la notion de récompense peut-être très difficile à estimer et à mesurer (Su et coll., 2016b; Li et coll., 2020b) notamment lorsque la tâche à effectuer est complexe.

Quelle que soit l'approche utilisée, la politique du dialogue peut-être mise à jour en utilisant le *feed-back* des utilisateurs (Su et coll., 2016a; Liu et coll., 2018). Selon le type d'apprentissage (supervisé, par renforcement ou adverse), il existe différentes stratégies d'apprentissage interactif. Dans cette thèse, nous avons prévu de suivre une approche itérative qui sera détaillée dans la section 3.3.4. Ainsi, durant le développement du système il est possible d'ajuster la politique au fur et à mesure des phases de modélisation et de développement.

2.4.6 Génération automatique de textes

Lorsque le système effectue une action, le retour à l'utilisateur, c'est-à-dire la réponse du système, doit être formulé en langue naturelle : tel est l'objectif de ce composant. Dans un dialogue humain, les réponses ne sont pas choisies à partir d'une liste prédéfinie de phrases. Cependant, dans une architecture simple, pour toute action, le système l'associe à un énoncé prédéfini sous forme d'une réponse à l'utilisateur. Dans les systèmes classiques, les énoncés prédéfinis font partie d'un schéma (*template* en anglais) capable d'utiliser les concepts remplis par le système comme des variables. Ces schémas peuvent comporter des règles de transformation comme dans le système de Weizenbaum (1966).

Afin de se rapprocher plus vers des dialogues humains, plusieurs possibilités de réponses prédéfinies à une action peuvent être établies. De même, la génération automatique de textes peut être réalisée en deux étapes : dans un premier temps, la planification du contenu basée sur les cadres puis la réalisation de surface qui nécessiterait un corpus adapté au domaine d'application. Les approches neuronales classiques utilisent des méthodes de séquences à séquences permettent d'effectuer une génération de texte à partir d'une représentation de sens afin d'éviter ce processus de génération en 2 parties (Puzikov et Gurevych, 2018). La Figure 2.9 montre un exemple de représentation de sens avec une phrase de référence associée.

MR:

name[The Eagle]	*eatType[coffee shop]*
food[French]	*priceRange[moderate]*
customerRating[3/5]	*area[riverside]*
kidsFriendly[yes]	*near[Burger King]*

Human Natural Language Reference:

"The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King."

FIGURE 2.9 – Exemple d'une représentation de sens et la génération d'une phrase par le module de NLG (Puzikov et Gurevych, 2018)

La GAT est étroitement liée à la théorie des actes de parole que nous avons abordée dans la section 2.2.1. La représentation de sens est donc modélisée comme des concepts et valeurs comme pour les systèmes de dialogue. Cette tâche de conversion d'une représentation de sens (dénoté comme *MR* sur la Figure 2.9 à une chaîne de caractères peut-être vue comme la tâche inverse de la NLU. La génération des énoncés dans un cadre dialogique avec les méthodes séquences à séquences est répandue dans le domaine (Wen et coll., 2015a,b; Shang et coll., 2015). Les systèmes séquences à séquences avec de l'attention sont performants dans la production des énoncés grammaticaux. En revanche, ils ont du mal à copier les détails, notamment les mots rares. Face à ce problème, un mécanisme de copie est couramment utilisé (See et coll., 2017; Gehrmann et coll., 2018). Dans un contexte dialogique, ces systèmes proposent souvent des énoncés qui sont génériques ou qui ne sont pas très originaux (Jiang et de Rijke, 2018) et peuvent être très répétitifs. Les paramètres de l'algorithme de recherche en faisceau utilisé souvent dans le décodage a un impact important sur les énoncés produits.

2.4.7 Synthèse vocale

Dans un système oral, le composant de synthèse vocale consiste à transformer le contenu textuel produit par le système de GAT à un signal de la parole. Les premiers systèmes de synthèse vocale consistaient à décomposer la phrase en petites composantes acoustiques qui étaient ensuite prononcées séquentiellement par le système.

Cependant, comme tous les autres sous-composants, les approches stochastiques sont de plus en plus utilisées dans le domaine de la synthèse vocale. Traditionnellement, il y avait deux approches majeures utilisées dans le domaine de la génération de la voix artificielle : approche par concaténation (Lee et Cox, 2002) et la synthétisation vocale à formants (Högberg, 1997). Comme son nom l'indique, l'approche concaténative consiste à mettre ensemble (concaténer) des segments de parole d'un locuteur, enregistré auparavant dans de bonnes conditions, afin de composer des segments plus grands pour reproduire de la parole. La synthèse à base de formants est une technique à base de règles permettant de simuler la structure de formant de langue naturelle.

L'une des approches neuronales couramment utilisées dans le domaine est celle basée sur l'architecture Tacotron2 (Shen et coll., 2018). Dans cette architecture l'encodeur encode une chaîne de caractères sous forme de vecteur de caractères convolutionnels. Le décodeur effectue la prédiction du cepstre et l'utilise pour générer du son avec *Wavenet* (Oord et coll., 2016). Dans notre système, nous avons opté pour la synthèse vocale utilisée par défaut sur les terminaux mobiles. Pour plus de détails sur l'état de l'art de la synthèse vocale, la revue de détaillé des méthodes neuronales est présenté récemment par Ning et coll. (2019).

2.5 Compréhension automatique du langage naturel

La première tâche d'un système de compréhension, orale ou textuelle, est de faire une classification de domaine afin de déterminer le champ sémantique. En effet, dans un système capable de réaliser plusieurs tâches prédéfinies, les tâches peuvent être intrinsèquement indépendantes les unes des autres : par exemple, un système pourrait être capable de gérer la prise de rendez-vous avec les cliniciens ainsi que d'obtenir des informations sur un médicament. Dans un tel système, les cadres représentant la prise de rendez-vous et ceux permettant d'obtenir des informations pharmacologiques seront conceptuellement différents.

Après avoir défini le domaine de l'énoncé, la seconde tâche de la NLU est d'identifier l'intention du locuteur et déterminer l'acte de parole de l'utilisateur. Une fois que le système a identifié le domaine d'application et l'intention de l'utilisateur, la tâche suivante est l'extraction d'informations pertinentes de l'énoncé afin de créer une représentation sémantique de la phrase. Le tableau 2.4 schématise la représentation sémantique extraite de la phrase.

Phrase	je	voudrais	savoir	les	précautions	d'emploi	du	Doliprane ©	500	mg
Attributs	O	O	O	O	B-request	I-request	O	B-drug	B-dos	B-unit
Intention	request_information									
Domaine	pharmacologie									

TABLE 2.4 – Représentation sémantique de l'énoncé du point de vue de la NLU

Sur le Tableau 2.4, la toute première information extraite concerne le domaine : la pharmacologie. L'intention de l'utilisateur est identifiée comme « obtenir une information ». Dans cet exemple, les mots de l'énoncé «je voudrais savoir» sont essentiels pour la classification, mais ne font pas partie des cadres définis pour cette tâche. Ensuite, l'information pertinente identifiée est celle qui est demandée par l'utilisateur : les précautions d'emploi et informations sur un médicament dont le nom commercial, son dosage et l'unité utilisée sont précisés par l'utilisateur. Dans la saisie d'attributs, nous distinguons les cadres, des valeurs qu'ils portent. Par exemple, le cadre *request* aurait pour valeur *précautions d'emploi*. Il est à noter que les valeurs portées par les concepts ne sont pas atomiques. Pour la valeur « mg » de l'entité 'B-unit' du tableau 2.4, nous pourrons imaginer d'autres valeurs comme milligramme,milligrammes,millig.,etc. Dans NLU, les valeurs capturées par les concepts ne sont pas forcément des synonymes. Ils peuvent être un mot écrit avec une erreur d'orthographe, une reformulation, un acronyme qui représente le concept, etc. Pour cette raison, on trouve souvent une couche supplémentaire des valeurs normalisées. Une entité peut être représentée avec ces trois informations :

— Slot-étiquettes : Le nom de l'entité qui définit le cadre sémantique. Sur le Tableau 2.4 ce sont les différents attributs tels que request, drug, dos ou unit. Les schéma d'annotations viennent s'ajouter à ces entités pour définir le début, la fin, etc. des annotations.

— Slot-valeurs : Par les slots valeurs, on représente les valeurs normalisées. Par exemple, dans le domaine médical Campillos et coll. (2018); Tourille et coll. (2017) utilisent les

concepts UMLS pour normaliser les termes médicaux en se basant sur les terminologies internationales du domaine.

— Valeurs : La valeur réelle (chaîne de caractères) du slot.

Les marqueurs (I,O,B) font partie de ce qui est appelé comme un schéma d'annotation. Dans la littérature il y a plusieurs schémas qui sont utilisés pour annoter les entités. Les plus classiques sont I=Inside(Dedans), O=Out(Dehors) et B=Begin(début). Le choix du schéma d'annotation est difficile à déterminer et dépend de la tâche (Konkol et Konopík, 2015) et la littérature du domaine. À part ces marqueurs classiques, il existe d'autres marqueurs tels que E=End(Fin), L=Last(Dernier), U=Unit(Unité). Ces différents marqueurs peuvent être combinés comme IO,IOB,BILOU,etc. La plus classique est celle qui est démontrée sur le tableau **??**, de marqueur le début, la continuité et les mots qui ne sont pas des entités(O).

Les premiers travaux sur la NLU ont commencé dans les années 70' en utilisant des techniques à base de connaissances (Wang et coll., 2011). Les techniques utilisées actuellement sont basées fondamentalement sur des approches à base de données. La fabrication de grammaires étant coûteuse et sujette à erreurs, les premiers systèmes ont introduit des approches statistiques génératives à base des modèles Markoviens. Puis, pour bénéficier du caractère linéaire de la langue, les systèmes se sont orientés vers des modèles discriminatifs : les champs aléatoires conditionnels (CRF). Les CRF ont été récemment dépassés par des modèles à base de réseaux de neurones : des réseaux de neurones récurrents (Mesnil et coll., 2014), Bi-LSTM encodeur-decodeurs (Bapna et coll., 2017) tels que le RNN à base de modèles d'attention (Liu et Lane, 2016) ou le CNN avec le mécanisme d'attention (Huang et coll., 2017).

Comme illustré sur le Tableau 2.4, l'approche la plus courante est de voir la saisie d'attributs comme un problème d'étiquetage où chaque mot de l'énoncé est associé à un attribut. Cependant, d'autres approches alternatives peuvent être envisagées pour gérer le problème de classification (Henderson et coll., 2012) comme le modèle seq2seq avec la génération de pointeurs ou étiquetage de dépendances (Huang et coll., 2017).

La tâche de la détection d'intention est traitée comme un problème de classification et est considérée comme un problème distinct de la saisie d'attributs. Cependant, comme les deux tâches sont fortement corrélées, certaines approches récentes ont abordé ces deux tâches simultanément. Par exemple l'apprentissage jointe de l'intention, slot-étiquettes et slot-valeurs à partir des données non annotées (Mishakova et coll., 2019a) ou le Tri-CRF (Jeong et Lee, 2008) qui étend la séquence linéaire étiquetant le CRF par un nœud pour représenter l'acte de dialogue, et Att-RNN (Liu et Lane, 2016) qui étend le RNN encodeur-décodeur avec un décodeur d'intention supplémentaire. Dans cette section nous allons décrire d'abord les approches classiques utilisées dans le domaine de NLU dans 2.5.1, puis détailler les approches neuronales dans 2.5.2.

2.5.1 Approches classiques

Les premières approches à être utilisées dans le domaine de la compréhension automatique du langage étaient à base de règles (Weizenbaum, 1966; Seneff, 1992; Wallace, 2009). Ces systèmes s'appuient majoritairement sur des automates à états finis et de la recherche de règles et de motifs. Comme présenté dans la section 2.3.1, malgré les bonnes performances par l'imitation, la construction de règles est longue et laborieuse. De plus, dans un système de dialogue oral, nous sommes confrontés aux erreurs de RAP, des particularités de la parole spontanée telles que la non-grammaticalité, des reprises, hésitations, pauses, disfluences, etc. La compréhension automatique basée sur la théorie de cadres utilise la méthode de remplissage d'attributs que nous avons détaillé dans la section précédente. L'une des premières approches utilisées consiste à traiter le problème de remplissage d'attributs comme une grammaire hors contexte (Ward, 1990). L'une des premières applications de SLU dans les années 90' a été Wang et Acero (2002) avec le parseur Phoenix entrainé sur les données ATIS. Plusieurs approches à base des grammaires hors-contexte ont été proposées dans la littérature comme les grammaires catégorielles combinatoires (Steedman, 1996) et l'algorithme de Cocke-Younger-Kasami (CYK) (Dunlop, 2014).

Le remplissage d'attributs peut-être vu aussi comme une tâche de classification. Par exemple, Wang1 et coll. (2002) propose une approche mixte utilisant des règles et des méthodes statistiques. De manière générale, nous pouvons classifier les modèles statistiques en deux catégories : génératifs et conditionnels. Les premiers travaux utilisent des méthodes génératives telles qu'un classifieur bayésien naïf, l'estimation du maximum de vraisemblance, machine à vecteur de support (SVM) et les n-grammes (Chelba et coll., 2003; Wang1 et coll., 2002; Yang et coll., 2017). Les SVM sont notamment utilisés dans la prédiction d'intention et sont considérés comme *baseline* dans certains *benchmarks* (Haffner et coll., 2003). Le remplissage d'attributs vu comme une tâche de classification porte aussi des désavantages. En premier, les données annotées peuvent être limitées ou très peu (scarce), ou non équilibrées.

Les approches statistiques nécessitent que les énoncés soient représentés par des représentations vectorielles telles que les sacs de mots ou les plongements lexicaux (*word embeddings* en anglais). Le modèle du sac de mots est la méthode de classification de texte la plus couramment utilisée, où la fréquence de l'occurrence de chaque mot est utilisée comme caractéristique pour l'apprentissage d'un classificateur. L'objectif de ces représentations est de caractériser les mots par leurs contextes (Mikolov et coll., 2013). Les plongements de mots nous permettent d'utiliser une représentation efficace et dense dans laquelle les mots similaires ont un encodage similaire. Pour ce faire, les plongements lexicaux utilisent une représentation vectorielle numérique du texte à partir d'un corpus qui associe chaque mot du vocabulaire à un ensemble de vecteurs à valeur réelle dans un espace prédéfini à N dimensions. Les plongements lexicaux tentent de capturer le sens sémantique, contextuel et syntaxique de chaque mot du vocabulaire en se basant sur l'utilisation de ces mots dans les phrases. Les mots qui ont une signification sémantique et contextuelle similaire ont également des représentations vectorielles similaires.

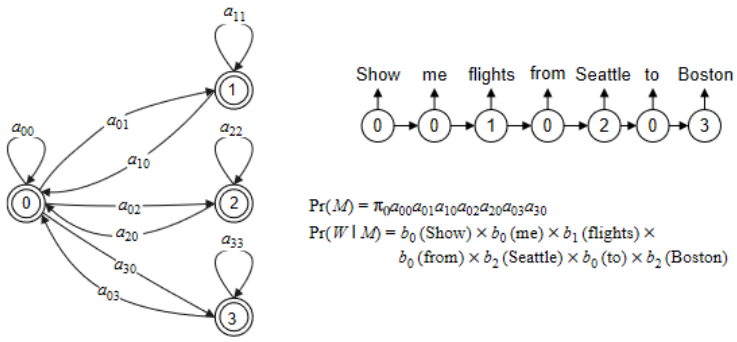


FIGURE 2.10 – Principe de fonctionnement du modèle de markov caché appliqué au domaine de la compréhension (Tur et De Mori, 2011)

Les modèles génératifs ont pour but de maximiser la probabilité jointe de $P(M,W)$ étant donné une suite de mots W et une représentation sémantique M. La probabilité jointe peut-être exprimée ainsi :

$$P(W,M) = P(W|M)P(M) \tag{2.3}$$

L'une des méthodes plus utilisées parmi les modèles génératifs est celle de Markov cachée (HMM). La Figure 2.10 illustre le principe de fonctionnement d'un HMM.

Dans le modèle HMM schématisé sur la Figure 2.10, les cadres sémantiques sont représentés par les états sur la gauche de l'image, marqués comme 1,2,3, auxquelles un sens est attribué. Les $a_{00} - a_n$ dénotent les transitions entre les états. Les $b_0 - b_n$ représentent les mots (observations) qui remplissent les cadres sémantiques (états cachés). Le sens d'une phrase est représenté par sa séquence d'états sous-jacente. Parmi les systèmes historiques basés sur les HMM, nous pouvons citer le système CHRONUS appris sur le corpus de ATIS(Levin et Pieraccini, 1995).

Les modèles de Markov cachés sont génératifs et produisent des résultats en modélisant la distribution de probabilité conjointe. Deuxième catégorie de modèles statistique est celle qui sont conditionnelles. Le modèle le plus utilisé dans le domaine de la compréhension a été les champs aléatoires conditionnels (*conditional random fields*- CRF). Les champs aléatoires conditionnels sont discriminants et modélisent la distribution de probabilité conditionnelle. Les CRF ne reposent pas sur l'hypothèse d'indépendance (les étiquettes sont indépendantes les unes des autres) et évitent ce biais d'étiquette. Les modèles HMM peuvent être vus comme un cas très spécifique de champs aléatoires conditionnels. La Figure 2.11 tirée de Jeong et Lee (2008) illustre les champs aléatoires conditionnels.

La Figure 2.11 présente les différents types de CRF. (A) représente les CRF à chaîne linéaire qui implémentent des dépendances séquentielles dans les prédictions. (B) représente les CRF factorielles (Sutton et coll., 2007). Les CRF factoriels sont des distributions conditionnelles qui ont plusieurs couches de chaînes linéaires avec des connexions entre les étiquettes. (C) représente les CRF avec des états cachés permettant de modéliser la structure latente de l'entrée avec une distribution conjointe sur l'étiquette de classe et les étiquettes

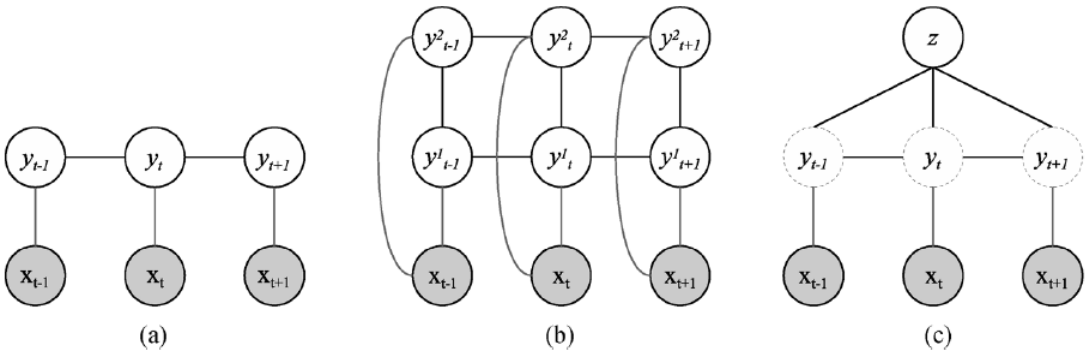


FIGURE 2.11 – Schématisation des champs aléatoires conditionnels (Jeong et Lee, 2008)

d'état cachées conditionnées par les observations, avec des dépendances entre les variables cachées exprimées par un graphe non orienté (Quattoni et coll., 2007). Dans le modèle CRF, les noeuds blancs représentent les variables aléatoires (le sens). Étant donné une séquence de mots (observation) dénotés comme x_{t-1} à x_{t+1}, pour un intervalle de temps $t-1$ à $t+1$, ce modèle a pour objectif de trouver la probabilité d'une séquence d'étiquettes y étant donné une entrée de vecteurs de séquence (X). Cette probabilité est désignée par $P(y|X)$. Dans l'approche classique basée sur une chaîne linéaire modélisant la distribution de probabilités conditionnelle de la séquence de sortie étant donné une séquence d'entrée. Dans CRF, l'étiquette y_t est conditionnellement dépendant sur l'étiquette y_{t-1}.

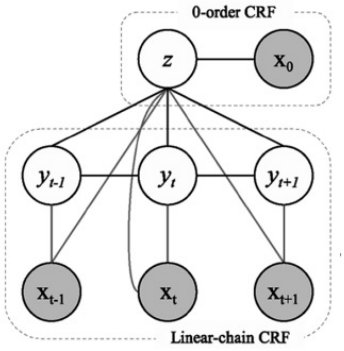


FIGURE 2.12 – Représentation graphique du modèle tri-crf (Jeong et Lee, 2008)

Le CRF est considéré comme *baseline* dans la plupart des tâches d'étiquetage de séquences et était considéré comme l'état de l'art avant les méthodes neuronales. L'un des avantages de l'approche CRF repose sur la fonction caractéristique qui peut utiliser toute la séquence de données d'entrée. En effet, la probabilité de l'étiquette du segment de données observés peut dépendre de tous les segments passés ou futurs. L'étiquetage des séquences est souvent associé à la classification des séquences et peut-être vu comme une tâche jointe. Le modèle tri-crf propose une extension à ce modèle en proposant une étiquette supplémentaire z permettant de créer un lien de dépendance entre cette variable aléatoire z représentant l'intention et la séquence d'étiquettes cachées (Jeong et Lee, 2008, 2009). Le modèle est schématisé sur la figure 2.12.

2.5.2 Approches neuronales

Le succès des réseaux de neurones dans d'autres domaines du TAL a suscité un intérêt pour le domaine de NLU notamment par l'introduction des réseaux de neurones récurrents (RNN). Les RNN sont des réseaux neuronaux dotés de boucles pour conserver les informations. Les RNN sont appelés récurrents car ils effectuent la même tâche pour chaque élément de la séquence et les éléments de sortie dépendent des éléments ou d'états précédents. Les premiers travaux concentrent sur le remplissage d'attributs et appliquent différents types de RNN appris sur le corpus ATIS (Mesnil et coll., 2013; Yao et coll., 2013, 2014b). Les premiers travaux sur les RNN se concentrent sur la résolution classique des problèmes d'apprentissage machine dans le domaine de TAL tel que le problème de disparition du gradient, ce qui signifie que les dépendances à long terme ne peuvent pas être apprises par le modèle. De cette raison, les réseaux de neurones récurrents à mémoire court terme et long terme (LSTM) ont commencé à être utilisés plus fréquemment en raison de leur capacité à oublier les informations et à modéliser avec plus de succès les dépendances à long terme (Yao et coll., 2014a; Peng et Yao, 2015; Kurata et coll., 2016).

Cependant, même les LSTM peuvent être moins performants avec des phrases très longues. Une approche qui répond à ce problème est l'utilisation des réseaux de neurones convolutionnels (Vu, 2016). En partant du même constat que le modèle Tri-CRF, Xu et Sarikaya (2013) proposent une adaptation du modèle Tri-CRF avec les CNN capable d'apprendre l'intention du locuteur ainsi que les cadres sémantiques jointement. En effet, la tâche du remplissage d'attributs et la détection d'intention sont fortement corrélées (Liu et Lane, 2016; Xu et Sarikaya, 2013). Dans cette optique, Liu et Lane (2016) propose une architecture encodeur-décodeur qui permet de faire la saisie d'attributs et la détection d'intention simultanément. L'encodeur du modèle Att-RNN est un LSTM RNN bi-directionnel qui prend en entrée les mots de l'énoncé encodant dans chaque cycle de temps un mot x_t. La sortie t de chaque cycle de temps correspond à l'état caché h_t du RNN bidirectionnel constitué de la concaténation de l'état caché de la propagation avant et arrière du RNN. Ainsi, l'état caché obtenu dans le dernier cycle de temps contient l'information sur toute la séquence d'entrée. Le fonctionnement du modèle Att-RNN est schématisé sur la Figure 2.13.

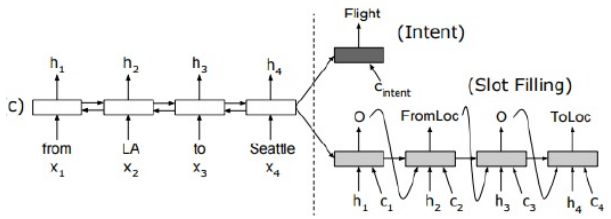


FIGURE 2.13 – Schéma du modèle Att-RNN (Liu et Lane, 2016)

Dans une architecture encodeur-décodeur, cet état caché est passé aux décodeurs pour initialiser leurs états cachés initaux. Comme dans le modèle tri-crf, l'un des décodeurs représente l'intention de la séquence d'entrée et attribue une intention correspondant à l'énoncé. Le deuxième décodeur est celui représentant les attributs qui produit une séquence d'attri-

buts correspondant à la taille du vecteur d'entrée. Le décodeur d'attributs est un LSTM RNN uni-directionnel qui produit une prédiction d'attribut dans chaque cycle du temps de décodage. Également, dans chaque cycle de temps comporte un vecteur de contexte d'attention c_t qui implémente le mécanisme d'attention (Bahdanau et coll., 2014).

Comme le système tri-crf, att-rnn effectue la prédiction de l'intention et des attributs mais pas les valeurs qui remplissent ces attributs. Contrairement au modèle CRF, dans ces systèmes, l'entraînement des données comportant des attributs et des données comportant les valeurs des attributs se fait séparément. Un exemple de représentation de données utilisé dans ces systèmes est schématisé dans le figure 3.6.

$$
\begin{aligned}
&medical\_prescription \\
&\qquad (inn = inn = Amoxicillin \\
&\qquad d-dos-val = d\_dos\_val\_numeric = 500 \\
&\qquad d-dos-up = milligram = mg \\
&\qquad dos-val = dos\_val\_numeric = 2 \\
&\qquad dos-uf = capsule = capsules \\
&\qquad rhythm-perday = rhythm\_perday = 2 \\
&\qquad dur-val = dur\_val\_numeric = 8 \\
&\qquad dur-ut = day = days)
\end{aligned}
$$

FIGURE 2.14 – Exemple d'attributs et valeurs alignés

Ces systèmes nécessitent une grande quantité de données annotées et alignées qui est dépendante du domaine. Cependant, il est possible également d'effectuer la tâche de compréhension du langage en utilisant un modèle séquence vers séquences (seq2seq) sans avoir besoin d'un alignement explicite (Mishakova et coll., 2019a). L'architecture de seq2seq est très similaire à Att-RNN, sauf qu'il ne comporte pas un classifieur pour l'intention. Dans un modèle seq2seq, l'intention est ajoutée à la prédiction comme un attribut en début de la séquence de sortie comme dans l'exemple ci-dessous :

```
intent [ prescription ], d-dos-up
[ milligram ], dos-uf [ capsule ]...
```

En termes de performances, même si les systèmes non alignés produisent des résultats de moins bonne qualité en général, ils permettent d'entraîner les systèmes utilisant des corpus de grande taille qui n'est pas annotée à l'échelle des mots.

Récemment, de plus en plus d'approches se basent sur l'architecture transformeur (Vaswani et coll., 2017) et utilisent notamment les modèles de langues pré-entrainés sur des corpus textuels de grande taille de façon non-supervisé. Les approches à base de transformeur sera détaille dans la section 2.8.

2.6 TAL biomédical - un domaine peu doté en France

Dans la section 2.1, nous avons abordé le domaine du TAL biomédical ainsi que les caractéristiques de la langue utilisée dans ce domaine. Dans le domaine biomédical, les résultats prometteurs des réseaux de neurones a attiré beaucoup d'attention dans la communauté ces dernières années (Wu et coll., 2020). En revanche, l'entraînement des modèles neuronaux nécessite une grande quantité de données. Contrairement aux autres domaines du TAL, lié au caractère médical des données, le partage des corpus médicaux (textuel, oral ou multimodal) est plus restreint, notamment pour les langues autres que l'anglais. Même si, les données sont peu accessibles par rapport aux autres domaines de TAL, il existe des entrepôts de données, des données ouvertes à la recherche scientifique à la demande et des données développés dans le cadre des défis.

Dans cette section, nous allons décrire d'abord les entrepôts de données utilisées couramment dans le domaine biomédical dans 2.6.1. Souvent associés à des données hospitalières, les entrepôts de données contiennent des données cliniques sous forme de base de données. De manière générale, ils sont utilisés dans la recherche médicale. Plus récemment, avec la disponibilité des données textuelles, il y a une contribution plus soudée entre le TAL et le domaine biomédical (Demner-Fushman et coll., 2009). Ensuite, nous allons détailler plus spécifiquement les défis organisés dans le domaine biomédical dans la sous-section 2.6.2 en mettant l'accent sur les défis sur l'extraction des informations liés aux médicaments. Dans ce domaine, la plupart des efforts sont concentrés sur l'extraction des informations cliniques à partir des textes médicaux. Nous allons brièvement présenter le domaine de l'extraction d'entités nommés dans le contexte médical dans la partie 2.6.3.

2.6.1 Entrepôts de données

Les institutions cliniques créent de plus en plus des bases de données qui archivent et organisent les données liées aux patients dans des « entrepôts de données ». Les données peuvent être issu des différents composants d'une institution, notamment la pharmacie, le laboratoire et la radiologie, etc. et peut contenir diverses informations liées au plan de soins cliniques (tels que les plans de soins infirmiers, les dossiers d'administration des médicaments et les ordonnances médicamenteuses).

Les entrepôts de données contiennent souvent des dossiers médicaux électroniques (Dossier Médical Partagé - DMP) des patients qui permettent d'archiver et d'organiser les dossiers des patients. Dépendant des pays et l'institution, la forme et le contenu des dossiers patients peuvent être différents. Le DMP contient une synthèse de l'historique médical des patients qui sont utilisés par les cliniciens qui font partie du plan de soins du patient. Le DMP contient des informations démographiques, antécédentes physiques, ordonnances, allergies, les notes des infirmières, etc.

L'entrepôt de données historique très fréquemment utilisé pour la recherche médicale est l'entrepôt de données MIMIC (Lee et coll., 2011; Johnson et coll., 2016, 2020). Dévelop-

pée dans le cadre d'une collaboration entre l'université de MIT et des partenaires privés en 2003, ce base de données est ouverte à la recherche gratuitement. À l'heure actuelle, il y a 4 versions du MIMIC. Les travaux sur l'apprentissage automatique se concentrent sur la base de données MIMIC-III. Il contient environ 50 tables comprenant diverses informations sur les données cliniques des patients telles que les examens médicaux, les données d'imagerie médicale, données liées à la durée d'hospitalisation, etc. Pour la recherche du TAL biomédical, cette base de données est particulièrement intéressante parce qu'il contient des comptes rendus de sortie d'hospitalisation. La plupart des applications focalisent sur les données structurées afin de construire des algorithmes d'aide à la décision médicale mais 80% des informations sur un EHR fait partie du texte de façon non structurée (Huang et coll., 2019). En France, l'un des entrepôts de données le plus connus est celui de l'hôpital AP-HP à Paris. Cependant, contrairement au MIMIC, l'accès aux données est conditionné par l'acceptation d'une commission qui évalue la faisabilité de la recherche, les nécessités réglementaires, l'aspect éthique et partenarial. D'autres hôpitaux contiennent des entrepôts de données qui fonctionnent d'une manière similaire comme l'entrepôt de l'hôpital de Lille ou de Rouen.

Pour la recherche médicale et innovation, diverses informations sont publiées périodiquement par les gouvernements ou les ministres de la santé. Par exemple, Open NHS (centre national de la santé) met à disposition des données ouvertes aux chercheurs qui est mis à jour périodiquement. De la même manière, en France, le portail des données publiques françaises met à disposition des informations incluant le domaine de la santé en France disponible gratuitement [4]. Ces données font partie des données structurées. Concernant les données non structurées, la littérature scientifique sur la santé est utilisée souvent sous forme de corpus. Cette initiative est menée par les plateformes d'archives ouvertes telles que NLM (*National Library of Medicine*) [5] pour le domaine biomédical, HAL en France, etc. Récemment, l'utilisation des données issues de Tweets (issu de la plateforme Twitter) est devenue courante (Aramaki et coll., 2011; Velardi et coll., 2014; Abboute et coll., 2014), notamment pendant l'épidémie de Covid-19 (Müller et coll., 2020).

2.6.2 Défis, collectes et annotations des données médicales

Dans le domaine biomédical, en plus du coût important de l'annotation manuelle proprement dite, il est nécessaire d'anonymiser (ou pseudo anonymiser) les données pour protéger la vie privée des patients et des soignants. En conséquence, les données recueillies sont généralement restreintes au projet de recherche et ne sont généralement pas diffusées. L'étude de Wu et coll. (2020) montre que la moitié des corpus présentés dans les articles scientifiques font partie des corpus institutionnels ou privés non disponibles à la communauté scientifique. Cela rend la reproductibilité de la recherche difficile (Chapman et coll., 2011; Collins et Tabak, 2014). Dans ce contexte, il y a un effort considérable dans la création de défis et la collecte de données qui vise à produire des outils, des données de référence ainsi que des données d'annotation qui sont ouvertes à la communauté scientifique.

4. https ://www.data.gouv.fr
5. https ://pubmed.ncbi.nlm.nih.gov/

Défi	Sujets
I2B2 : Informatics for Integrating Biology and the Bedside	2006- Identification des fumeurs, 2008-Détection de l'obésité, 2009 - informations sur les médicaments, 2010- relations, 2011- coréférences, 2012- relations temporelles, 2014- identification des maldies du coeur
N2C2 : National NLP Clinical Challenges	2018- les études de cohort, 2018- extraction de l'information et détection des EIM sur les médicaments
CLEF-eHEALTH	Lab d'évaluation 2013 - 2019 : l'extraction d'information, gestion de l'information et la recherche d'information. Pour une revue détaillé consultez leur site officiel*
eHealth-KD	2019&2020- annotation sémantique (espagnol)
The China Conference on Knowledge Graph and Semantic Computing (CCKS)	2017-2020 : extraction d'entités nommés et les événements (chinois)
SemEval	2013- Extraction d'interactions entre les médicaments, 2014-15 : Extraction d'entités nommées et l'annotation des concepts 2016-17 : Extractions des relations temporelles 2018 : Extraction des relations sémantiques
Centers of Excellence in Genomic Science (CEGS) N-Grid	2016- Extraction de la sévérité des symptômes
Medication and Adverse Drug Events (MADE)	2019- Extraction des EIM, indications et informations sur les médicaments
MedNLPDoc	2016- Extraction d'information (japonais)

TABLE 2.5 – Synthèse des défis du domaine du TAL biomédical (*:https://clefehealth.imag.fr/)

L'une des premières initiatives dans le domaine du TAL biomédical sur le développement de ressources est les compétitions de I2B2 (*Informatics for Integrating Biology and the Bedside* en anglais) [6]. Le premier défi de cette plateforme a été sur la désidentification des comptes rendus médicaux et la détection et l'identification de fumeurs ces derniers (Uzuner et coll., 2006) en 2006. Depuis plusieurs défis ont été organisés : l'extraction d'information liée à l'administration de médicaments (Uzuner et coll., 2010a), la relation entre les entités (Sun et coll., 2013), la résolution de coréférences (Uzuner et coll., 2012), etc. Depuis 2018, le défi N2C2 (*National NLP Clinical Challenges*) a pris le relais d'I2B2 et continue sur la même voie qui est organisée sous forme de *workshop* comme les défis d'I2B2. Les ressources que nous avons abordées sont en anglais, mais il y a des initiatives dans la création des données annotées dans d'autres langues que l'anglais. Par exemple, nous pouvons citer le défi CLEF-eHEALTH qui a organisé plusieurs campagnes de compétitions sur 3 thématiques : l'extraction des entités nommées, la gestion de l'information et la visualisation et la recherche d'information. En 2015, ils ont organisé une compétition sur la reconnaissance d'entités nommées pour le français sur le corpus QUAERO (Névéol et coll., 2014). Depuis, il y a eu d'autres campagnes telles que l'annotation automatique des codes ICD-10 sur les textes biomédicaux en allemand (Kelly et coll., 2019), en espagnol (Piad-Morffis et coll., 2019; Miranda-Escalada et coll., 2020), en français (Névéol et coll., 2016) et en japonais (Aramaki et coll., 2016). Le tableau 2.5 présente une synthèse des défis dans le domaine de TAL biomédical.

6. https://www.i2b2.org/NLP/DataSets/Main.php

2.6.3 Reconnaissance d'entités nommées biomédicales

Comme son nom l'indique, la reconnaissance d'entités nommées (*Named Entity Recognition* ou NER) est une tâche qui consiste à reconnaître des entités à partir du texte. Cette tâche est très proche de celle de l'extraction de cadres de la saisie d'attributs abordée dans la section 2.5. Historiquement, la tâche de NER consistait à annoter les personnes, organisations et les lieux. L'application de NER dans le domaine biomédical (bio-NER), aurait pour but d'annoter les entités médicales : ARN, protéines, types de cellules, ADN, médicaments, maladies, symptômes, etc. Les entités concernées dépendant donc de la tâche. La Figure 2.15 montre un exemple d'annotation sémantique sur un document issu du défi de l'extraction de l'information sur les médicaments d'I2B2-2010 (Li et coll., 2010).

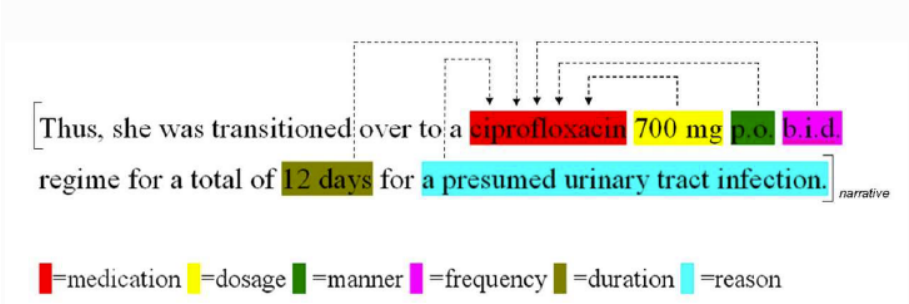


FIGURE 2.15 – Exemple d'annotation d'entités nommées médicales sur le corpus I2B2 (Li et coll., 2010)

Sur la Figure 2.15, on voit la visualisation des entités nommées annotée d'une phrase issu du défi I2B2-2010. Dans ce défi, les entités nommées sont : le médicament (rouge), dosage du médicament (jaune), la voie d'administration (vert clair), la fréquence de la prise (rose), durée de l'administration du médicament (vert foncé) et la raison d'administration du médicament (bleu). À la différence de NLU, l'annotation ne s'effectue pas sur des énoncés mais sur des textes souvent longs. Dans cet extrait, la raison de la prise du médicament est exprimée sur la même phrase, mais elle aurait pu être exprimée bien avant dans le document, ou après. De la même manière, une raison pourrait s'appliquer à la prise de plusieurs médicaments. La visualisation de la Figure 2.15 illustre cela avec les flèches qui montrent la relation entre ces entités.

Historiquement, les approches classiques utilisées dans le TAL biomédical étaient à base de règles, tels que MetaMap (Aronson et Lang, 2010), Medlee (Friedman, 1997) et KnowledgeMap (Denny et coll., 2003). Les premiers défis d'I2B2 ont montré que les approches à base d'apprentissage machine telles que les SVM et les CRF ont montré des résultats prometteurs (De Bruijn et coll., 2011; Tang et coll., 2013). Certaines approches ont abordé ce problème d'étiquetage de séquence avec des approches hybrides en s'appuyant sur MetaMap ou cTakes (Jiang et coll., 2011). Comme dans le domaine de NLU, les approches à base d'apprentissage machine ont certains inconvénients tels que l'ingénierie des caractéristiques assez sophistiques, gestion des dépendances au long terme ou encore le besoin de représentation de caractéristiques plus robuste que l'approche de sac de mots (Wu et coll., 2017).

Parmi les approches d'apprentissage machine appliquées dans ce domaine, le plus courant est d'utiliser un RNN bi-directionnel de type LSTM avec une couche de CRF (Jagannatha et Yu, 2016; Huang et coll., 2015) ou avec une couche de CNN (Chiu et Nichols, 2016) utilisée dans la génération des étiquettes. Mais d'autres approches existent comme l'utilisation d'une structure arborescente (Dinarelli et Rosset, 2011). Les approches récentes s'appuient souvent sur les plongements lexicaux contextualisés et utilise l'architecture transformeur (Vaswani et coll., 2017) tels que le modèle BERT (Li et coll., 2019; Yan et coll., 2019). Pour une revue plus détaille des approches utilisées dans le domaine de NER, les lecteurs peuvent consulter la revue de Li et coll. (2020a) et pour une revue des méthodes appliquées dans le domaine du TAL biomédical la revue de Wu et coll. (2017).

2.7 Méthodes d'augmentation de données dans le domaine biomédical

Malgré un effort considérable de création de ressources dans le domaine biomédical avec les défis, et les tâches partagées présentés dans 2.6.2, les données d'entraînement sont limitées. Ce manque de ressources est plus flagrant pour d'autres langues que l'anglais (Névéol et coll., 2018). Dans la littérature il y a eu diverses approches utilisées pour remédier à ce problème. Dans cette section, nous allons brièvement présenter ces approches. La partie 2.7.1 présentera les approches basées sur des automates à état fini utilisées comme une grammaire de génération pour la génération de données. La partie 2.7.2 présentera l'une des approches courantes utilisées pour augmenter les données dans d'autres langues que l'anglais par le biais de la traduction de ressources. Enfin, la partie 2.7.3 présentera les approches qui se basent sur un apprentissage semi-supervisé.

2.7.1 Génération des données artificielles par une grammaire

L'une des premières approches utilisées dans le TAL pour la génération des données synthétiques a été par de manière déclarative par l'utilisation du langage de programmation Prolog (Gal et coll., 1991; Dougherty, 2013). Prolog est un langage de programmation logique interprété dont le compilateur transforme les expressions logiques du premier ordre en séquence d'instructions. Plus tard, l'utilisation des grammaires hors-contexte (CFG) a été répandue dans l'informatique. Le principe de fonctionnement d'une grammaire est détaillé dans la partie 2.5.1. Avec une grammaire, nous décrivons un langage S avec les règles de haut niveau, des règles intermédiaires et des terminaux. Les CFG sont souvent utilisées pour les tâches de repérage (parsing) ou d'étiquetage (tagging). En même temps, une fois que les terminaux d'une grammaire est défini, le langage engendré par la grammaire S serait capable de produire toutes les chaînes acceptées par les définitions de la grammaire (Bird, 2006). Cette méthode pourrait être vue comme la tâche inverse de repérage, donc la génération. La génération artificielle de données a été appliquée dans le domaine biomédical pour générer des données de test randomisé (Maurer, 1990) et dans l'induction d'une grammaire hors-

contexte à partir de séquence d'ADN (Javed et coll., 2004). La génération du texte à partir d'une grammaire hors contexte (CFG) a été appliquée dans diverses tâches du TAL (Wong et Mooney, 2007; Le, 2007; Kusner et Hernández-Lobato, 2016). Par exemple, le système de Le (2007) (TGEN) utilise les CFG pour générer du texte dans une approche de remplissage d'attributs.

Avec la disponibilité des données, il y a de plus en plus une tendance d'utilisation des méthodes stochastiques dans la génération de données artificielles. Spécifiquement pour la génération de texte, l'approche la plus connue par le grand public est le modèle de langue GPT2/3 (Radford et coll., 2019; Brown et coll., 2020). Les modèles GPT-2/3 s'appuient sur l'architecture transformeur qui ont relevé le défi de la génération du texte avec de nombreux modèles appris sur un grand nombre de textes bruts. Parmi les modèles génératifs neuronaux, le modèle GPT-2 reste l'un des meilleurs pour la génération de texte. Des méthodes génératives telles que GPT-2 ont été récemment utilisés dans l'augmentation de données dans le domaine du TAL biomédical (Sybrandt et Safro, 2020; Amin-Nejad et coll., 2020) et par des méthodes seq2seq (Lee et coll., 2019). Une autre méthode neuronale à la génération de données consiste à utiliser les réseaux antagonistes génératifs (Goodfellow et coll., 2020) (GAN). Le GAN est un modèle génératif composé de deux réseaux en compétition avec l'un l'autre. Dans ce modèle, il y a un générateur et un discriminateur. Pour une revue détaillée de l'application des GAN dans le domaine biomédical, les lecteurs sont invités au consulter la revue récente publiée par Bissoto et coll. (2021).

Les approches génératives notamment à base d'architecture transformeur produisent des phrases de bonne qualité mais nécessite l'apprentissage sur un ensemble important de données de départ. Les méthodes à base de règles comme la grammaire de génération proposent une alternative permettant de générer des données de départ. Par exemple, Kusner et Hernández-Lobato (2016) utilise un CFG pour générer des données pour démarrer l'apprentissage d'un modèle GAN.

2.7.2 Approches à base de traduction

Dans le domaine biomédical, il y a de plus en plus de ressources en anglais, cependant pour d'autres langues que l'anglais, la disponibilité de ressources diminue largement. En effet, cela est confirmé par la revue systémique de Wu et coll. (2020) qui montre que plus de 70% des données de référence sont en anglais. Face à ce manque de ressources, l'une des premières méthodes appliquées dans le domaine était simplement de traduire les ressources. La disponibilité des services de traduction et l'augmentation de la qualité de traduction ont facilité cette approche. Les premiers travaux à appliquer une approche à base de traduction ont eu lieu bien avant la disponibilité de ces services tels que Claveau et Zweigenbaum (2005) qui propose une méthode de traduction automatique en utilisant les transducteurs pour la paire de langues anglais-français et Schulz et coll. (2004) depuis le portugais vers l'espagnol.

Cette méthode de traduction a été appliquée pour la traduction des comptes rendus médicaux (Liu et Cai, 2015), traduction de la parole dans le domaine médical avec des

tablettes (Rayner et coll., 2008), traductions de revues systématiques en français (Névéol et coll., 2013). La traduction automatique dans différents paires de langues a été également appliquée pour la recherche d'information multilingue (Névéol et coll., 2018).

Plus récemment, le domaine de la traduction automatique a vu un essor après l'introduction des méthodes séquences à séquences (Bahdanau et coll., 2014) puis le mécanisme d'attention sur soi avec l'architecture transformeur (Vaswani et coll., 2017). Également, des approches neuronales ont été appliquées dans la traduction des ressources médicales (Tubay et Costa-Jussa, 2018; Abdul-Rauf et coll., 2019). En dehors des approches neuronales, il existe d'autres initiatives qui visent à utiliser des ressources multilingues qui permettrait de transférer de la connaissance pour améliorer les performances des différentes tâches dans le domaine biomédical (Névéol et coll., 2018).

2.7.3 Apprentissage semi-supervisé

Les approches supervisées nécessitent des données d'apprentissage qui sont étiquetées souvent par l'effort humain. Dans un domaine complexe tel que le domaine médical, l'étiquetage nécessite une expertise médicale et donc par nature coûteuse. Malgré les ressources disponibles grâce aux défis et les tâches partagées, les ressources produites sont souvent utilisées comme des données de référence (test) pour estimer la performance des systèmes. Par exemple, dans le défi de l'extraction d'informations sur les médicaments d'I2B2-2010, l'équipe de Patrick et Li (2010) produit des données d'entraînement de seulement 145 documents sur 700, et gagnent le défi. Face à cette difficulté de ressources limitées, il y a un vrai intérêt dans le transfert de connaissance entre les différentes ressources médicales. Comme dans tout domaine d'apprentissage automatique, l'apprentissage supervisé est l'une des approches utilisées pour bénéficier de l'existence des différentes ressources pour améliorer la qualité d'apprentissage par transfert. Contrairement à une approche complètement non supervisée, dans l'approche semi-supervisée, nous disposons d'un ensemble de données annoté et l'objectif est de s'en servir des données non étiquetées pour améliorer la qualité de l'apprentissage. Cependant, la semi-supervision ne se limite pas seulement à l'utilisation des données non étiquetées mais peut concerner l'utilisation de règles, dictionnaires ou toute autre ressource qui permet le transfert de connaissance (Zhu et Goldberg, 2009).

Des approches semi-supervisées ont été appliquées dans diverses tâches des défis biomédicaux tels que le monitorat des symptômes de la dépression dans les réseaux sociaux (Yazdavar et coll., 2017), l'extraction d'informations sur les prescriptions avec un système à base de CRF (Tao et coll., 2018) ou le phénontypage à partir des comptes rendus avec un système basé sur l'algorithme de la maximisation de l'espérance (Dligach et coll., 2015). L'approche semi-supervisée est également appliquée aux approches neuronales. Par exemple Qader et coll. (2019) propose une approche semi-supervisée pour la génération du texte basée sur une architecture encodeur-décodeur. Pour la compréhension du langage, Qiu et coll. (2019) propose une architecture semi-supervisée à base graphes et l'appliquent à la tâche de détection de paraphrases.

Dans le domaine biomédical, malgré les très bonnes performances des systèmes utilisant les approches à base d'attention sur soi avec les plongements de mots contextualisés, certains systèmes semi-supervisés gardent leur place dans les défis. Introduit après la finalisation du défi, le système de Tao et coll. (2018) reste le meilleur sur la prédiction de la plupart des informations sur les prescriptions à partir des comptes rendus médicaux. Sur le défi d'I2b2-2010 sur l'extraction de relations, le système semi-supervisé de de Bruijn et coll. (2010) reste très compétitif.

2.8 Plongements lexicaux pré-entrainés dans le domaine médical

Le domaine du TAL a vécu un changement de paradigme après le succès des modèles BERT (les représentations d'encodeurs bidirectionnels) qui peuvent être appris sur un ensemble de texte de façon non supervisée. Le principe des modèles BERT se base sur l'architecture transformeur proposé par Vaswani et coll. (2017) qui porte un certain nombre d'avantages sur les approches séquentielles telles que LSTM, RNN, etc. La Figure 2.16 tirée de Vaswani et coll. (2017) présente l'architecture transformeur.

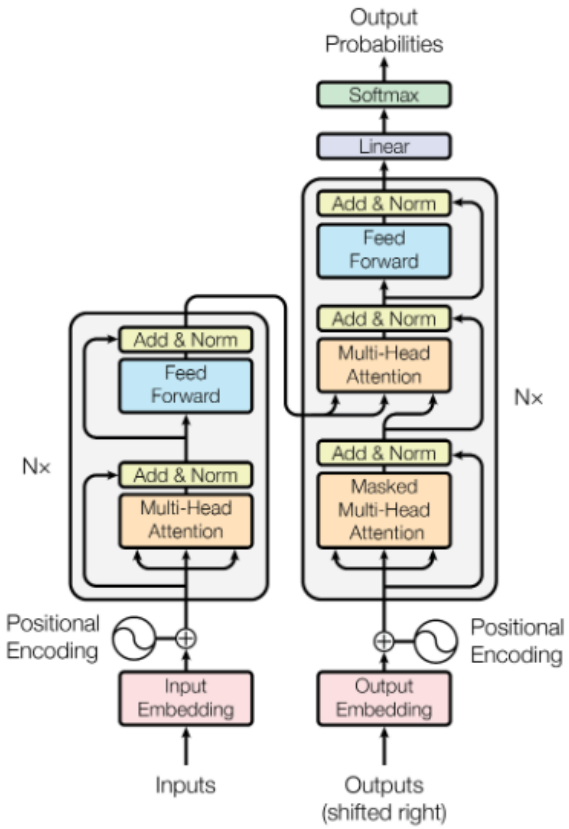


FIGURE 2.16 – Schéma de l'architecture transformeur (Vaswani et coll., 2017)

La Figure 2.16 illustre le schéma de l'architecture transformeur qui est basé sur une architecture encodeur et décodeur. Sur la Figure 2.16 la partie gauche représente l'encodeur et la partie droite représente le décodeur. Comme le nom de l'article l'indique, l'intuition derrière

cette architecture est celle que la seule attention est suffisante pour prédire une séquence de sortie. À la différence des approches séquentielles, dans l'architecture transformeur, toute la séquence d'entrée est encodée dans son ensemble par l'encodeur. Cette séquence encodée est passée à la couche d'attention qui calcule l'attention portée par les mots entre eux. L'encodage positionnel qui figure en bas à gauche permet de capturer cette similarité entre les mots de façon contextuelle en prenant en compte leurs positions.

Ce principe s'appuie sur le mécanisme d'attention. Il existe différents types d'attentions telles que l'attention concaténative (Bahdanau et coll., 2014) ou par produit scalaire puis concaténé et transmis par une couche linéaire finale. Le mécanisme d'attention utilisée dans transformeur est appelé comme le auto-attention ou attention sur soi (self attention mechanism) composé de plusieurs couches. Dépendant du mot sur lequel on paie de l'attention, les vecteurs d'attentions sont calculés dynamiquement. Ces différentes attentions sont calculées par les différentes têtes d'attentions. Vaswani et coll. (2017) présente deux modèles : un calculé avec 12 têtes d'attentions et un autre avec 24 têtes d'attentions qui correspondent en général aux plongements plus ou moins gourmands en termes de ressources. Comme dans une architecture classique de séquence vers séquences, la sortie des couches d'attention est passée par un réseau de neurones à propagation avant et est normalisée avant d'être transférée au décodeur. Dans le décodage, dans chaque cycle de temps t, le système prédit une séquence de sortie jusqu'à ce que le système arrive à une sortie spéciale qui dénote la fin de la séquence. Le mécanisme d'attention fonctionne de la même manière que dans l'encodeur et la position des séquences prédites est déterminée par une couche d'encodage positionnel. À la différence de l'encodeur, dans le décodage, l'auto attention paie de l'attention uniquement sur les séquences du cycle de temps t ou avant. Les séquences futures à prédire sont donc masquées. Les couches d'attentions sortent des probabilités qui sont transformées par la dernière couche de softmax en probabilités correspondant aux mots du vocabulaire.

L'apprentissage de façon non supervisé des corpus de textes a permis de produire des plongements lexicaux qui améliorent les performances des diverses tâches du TAL (Vaswani et coll., 2017). Dans cette section, nous allons brièvement présenter les plongements qui sont entraînés sur le français dans 2.8.1 et présenter de manière générale les plongements lexicaux utilisés dans le domaine de spécialité, notamment dans le domaine biomédical dans 2.8.2.

2.8.1 Plongements lexicaux pré-entrainés pour le français

Les représentations contextuelles apprises en utilisant l'architecture transformeur sur une grande quantité de textes produites des plongements lexicaux qui sont appelés comme des modèles de langue contextualisés pré-entrainés. Ensuite, sur une tâche du TAL, au lieu d'apprendre des représentations de mots à partir de zéro, ces représentations contextuelles sont chargées pour suivre un entraînement de réglage fin (*fine-tuning*). De cette manière, on permet un transfert de connaissance notamment utile lorsque les données d'entraîne-

ment sont limitées (Grießhaber et coll., 2020). En fait, cette méthode d'apprentissage est bien connue dans le domaine de l'apprentissage machine mais était plus récemment appliqué au TAL (Howard et Ruder, 2018).

Les modèles de langues pré-entrainés peuvent être répartis en 3 catégories : modèles d'encodage, modèles de décodage (méthodes génératives) et des modèles séquences à séquences. Le premier modèle de langue qui a été utilisé de cette manière est un modèle d'encodage : BERT (Devlin et coll., 2018) qui a été suivi par un mouvement des modèles de langues qui finissent avec « BERT ». L'entraînement du modèle BERT est effectué sur un ensemble de données constitué des articles de Wikipédia (2,5M de mots) ainsi qu'un ensemble de livres (800M de mots). L'apprentissage est effectué sur 2 tâches : Modélisation de langue masqué (Masqued Language Modeling- MLM) qui consiste à prédire le mot suivant étant donnée le début de la phrase et la prédiction de la phrase suivante (Next Sentence Prediction - NSP qui consiste à prédire si oui ou non, une certaine séquence A est suivie par une certaine séquence B. Pour prendre en compte d'autres langues que l'anglais, les modèles BERT incluent un modèle multilingue appris sur des articles de Wikipédia de 104 langues.

Bien que les modèles multilingues donnent des résultats intéressants sur d'autres langues que l'anglais, ils sont souvent plus volumineux et leurs résultats sont quand même moins bien comparés aux modèles monolingues (Martin et coll., 2019). Suivant ce constat, les premiers modèles de langue entraînés sur des corpus de textes en français sont *CamemBERT* (Martin et coll., 2019) et *FlauBERT* (Le et coll., 2019, 2020). Les deux modèles de langues se basent sur l'architecture *RoBERTa*(Liu et coll., 2019) qui est un dérivé de BERT. Les modèles de *RoBERTa* utilisent la même architecture que *BERT* mais suppriment la tâche de NSP, augmentent la taille de batches et entraînent le modèle pendant plus longtemps en effectuent des changements mineurs dans les paramètres qui donne de meilleurs résultats que les modèles initiaux. Pour le français, même si *FlauBERT* est appris sur un ensemble de données plus petit que *camemBERT*, il donne de meilleurs résultats que *camemBERT* sur certaines tâches.

2.8.2 Apprentissage par transfert de domaine

L'intuition derrière les modèles *BERT* repose sur l'exploitation et le transfert de connaissances en apprenant des représentations contextuelles avant d'apprendre des nouveaux domaines et tâches. Dans ces modèles de langues, on apprend « la langue » avant d'apprendre une tâche spécifique. Dans un domaine de spécialité tel que le biomédical, ce même principe est appliqué mais en apprenant uniquement sur des données issues de la langue de spécialité. L'un des premiers modèles de langue issue d'un domaine de spécialité était le *SciBERT* (Beltagy et coll., 2019) qui spécialise sur les articles scientifiques. Plus tard, les modèles de langues spécifiques ont été introduits dans plusieurs domaines, comme le domaine juridique (Chalkidis et coll., 2020), le domaine biomédical (Lee et coll., 2020), le domaine de tweets (Qudar et Mago, 2020), etc. Les résultats dans le domaine biomédical montrent que les modèles entraînés sur des corpus spécialisés augmentent largement les résultats des

différentes tâches biomédicaux par rapport à des modèles de langues généraux (Alsentzer et coll., 2019; Si et coll., 2019).

Avec les modèles *BERT*, pour adapter au domaine de spécialité, la technique utilisée par défaut est de reprendre l'entraînement du modèle sur un corpus de spécialité. Cela veut dire que l'entraînement reprend le modèle entrainé dans le domaine général puis re-entraîné sur des textes du domaine spécifique. Cependant, il est également possible de ré-entraîner le modèle *BERT* uniquement à partir des textes de spécialité. Par exemple, Alsentzer et coll. (2019) initialise l'apprentissage des modèles de langue biomédicale à partir de *BERT* et *bio-BERT*. Cependant, il est possible d'utiliser un corpus de spécialité au lieu d'initialiser le modèle de langue à partir du domaine général. L'un des plus récents et performants modèles de langues biomédicales proposées par Peng et coll. (2019) présente *BlueBERT* entraînés uniquement sur les notes cliniques du MIMIC-III et des résumés des articles scientifiques issus du PubMed [7]. Sur la tâche d'extraction d'information sur le médicament d'I2B2-2010, Dans Kocabiyikoglu et coll. (2021) nous montrons qu'un modèle basé sur BlueBert peut atteindre des performances à l'état de l'art.

Cette question de ré-entrainer ou entraîner soi même a été examinée plus en détail par El Boukkouri (2020). Son analyse systématique montre que les différentes configurations d'apprentissage ou de ré-apprentissage aboutissent à des résultats très similaires (El Boukkouri, 2020). Concernant les modèles de langues dans le domaine biomédical après bioBERT, il y a eu beaucoup de modèles de langue introduits dans le domaine biomédical. Une revue récente proposée par Kalyan et coll. (2021) détaille les caractéristiques et les performances des différents modèles de langues utilisées.

2.9 Conclusion sur l'état de l'art

Dans cet état de l'art, nous avons brièvement abordé les différents domaines de recherche liés à cette thèse : le TALN biomédical, les systèmes de dialogue et de compréhension automatique.

L'état de l'art sur les familles d'architectures des systèmes de dialogue présentés dans 2.4 nous montre qu'il y a principalement trois catégories : les systèmes à base de règles, les systèmes modulaires et les systèmes de bout en bout. Les systèmes de bout en bout ont l'avantage d'être plus facilement déployables parce qu'ils apprennent une représentation de la sémantique de la tâche à partir des données plutôt que de la représenter explicitement. Cette abstraction permet aux systèmes de choisir une réponse utilisateur sans rentrer explicitement dans la sémantique, en utilisant un corpus de dialogues. En revanche, dans le domaine médical, cette représentation est primordiale et nécessite l'utilisation et la vérification des informations comme le fait un LAP. Également, l'entraînement de ces systèmes nécessite une quantité de données importante et le système devrait être capable de s'adapter très rapidement aux changements sur les réglementations telles que l'autorisation de mise sur le marché (AMM).

7. https://pubmed.ncbi.nlm.nih.gov/

Pour ces raisons, nous avons choisi une approche modulaire (Williams et coll., 2016) dans la mise en place du système de dialogue. Cette méthode permettrait l'utilisation des approches neuronales récentes dans chacun des modules tout en bénéficiant d'un éventuel mécanisme de contrôle qui serait, par exemple, issu d'un LAP.

Dans une architecture modulaire, l'un des composants les plus importants est celui de compréhension. L'état de l'art présenté dans 2.5 nous montre que les systèmes les plus performants sont à base de réseaux de neurones et notamment ceux basés sur l'architecture transformeur qui a vu essor récemment. Cependant, l'un des facteurs à considérer est la quantité de données d'apprentissage qui est importante sur l'entraînement des modèles. La section 2.6 présente les corpus et les défis connus dans le domaine du TAL biomédical. Cette partie nous montre que la disponibilité des données, notamment pour le français, est assez faible. Il existe des méthodes pour gérer les cas où les données sont limitées présentées dans la section 2.8.2. Parmi ces approches, la génération de données semble intéressante dans notre cas. En effet, les prescriptions, même si elles peuvent contenir une écriture assez variée en fonction des différents cas de prescription, sont généralement composées d'expressions figées telles que le nom d'un produit, une durée, un dosage, etc. Ces données générées peuvent être utilisées pour amorcer le système de compréhension avant qu'il y ait une interaction réelle avec les utilisateurs. Puis, les méthodes semi-supervisées présentées dans 2.7.3 semblent intéressantes pour permettre l'utilisation des corpus médicaux issus des défis partagés dans l'amélioration des modèles de compréhension. Enfin, il est important de noter que les modèles de langues contextualisés ajustés à une tâche par *fine-tuning* semble être devenu la méthode la plus performante. Cependant, cette méthode n'a pas encore été testée sur des prescriptions médicamenteuses et nous contribuerons au domaine en évaluant une telle approche pour le français et l'anglais.

CHAPITRE 3

Méthode

Dans ce chapitre nous présenterons la problématique et la démarche de la thèse regroupées en plusieurs ainsi que les questions de recherche.

3.1 Circuit des prescriptions médicamenteuses

La saisie des ordonnances médicales pourrait-être définie comme un acte permettant aux professionnels de santé ayant l'habilitation, de prescrire un ou plusieurs traitements sur un document structuré qu'on appelle « l'ordonnance ». L'ordonnance consigne la prescription médicale qui est un acte qui se situe entre un conseil et un ordre, à la fois très structuré et conventionné par les lois, mais aussi très diverse et variée en termes d'information et d'instruction. Le prescripteur, celui qui instruit l'acte médical peut prescrire des médicaments mais aussi des examens complémentaires, des soins infirmiers, un régime, un transport médicalisé et même un arrêt de travail. Les médecins ne sont pas les seuls prescripteurs, les professionnels de santé ayant le droit de prescription sont nombreux : médecins, chirurgiens-dentistes, sage-femmes, infirmiers, masseurs-kinésithérapeutes, etc. Tous les professionnels de santé n'ont pas les mêmes droits, par exemple un infirmier peut prescrire des pansements mais ne pourra pas prescrire des médicaments comme le font les médecins. Par ailleurs, certains médicaments onéreux ou ceux sous surveillance renforcée nécessitent des prescriptions spéciales qui font l'objet d'un plan de gestion de risques. La prescription fait donc partie d'un circuit et d'une logistique complexe et permet la communication entre les professionnels de santé, d'où la nécessité qu'elle soit saisie sans ambiguïté. Cette communication met également en responsabilité les membres du corps médical (radiologues, biologistes, pharmaciens) concernant la validité de la prescription. L'analyse pharmaceutique de l'ordonnance consiste à calculer les doses prescrites, effectuer la comptabilité des solutés de perfusion, vérifier les interactions médicamenteuses et parfois à calculer les doses cumulées (Buonsignori, 2003). En effet, une partie des erreurs iatrogènes liées à la prescription de médicaments sont évitées dans le cadre de l'analyse pharmaceutique (Augry et coll., 1998; Schmitt, 2002). Les professionnels ayant le but d'administrer les médicaments, les auxiliaires médicaux (ex. infirmiers) n'ont pas la responsabilité de la validité de la prescription.

Pour assister le prescripteur dans cette saisie réglementée, les logiciels d'aide à la prescription proposent des interfaces permettant d'effectuer cette saisie de façon électronique. Cette saisie informatique permet de sécuriser la prescription en signalant les interactions

médicamenteuses potentiellement dangereuses, d'obtenir une meilleure gestion des stocks au sein de l'hôpital et permet d'avoir une intégration complète avec le dossier électronique des patients. En revanche, l'utilisation de ces interfaces, souvent complexes, nécessite un temps d'apprentissage préalable non négligeable et ces interfaces ne sont souvent accessibles que via un poste fixe.

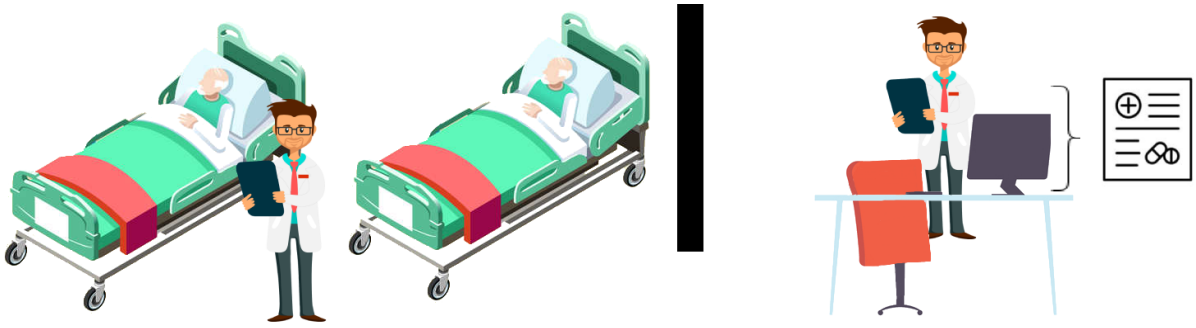


FIGURE 3.1 – Processus de diagnostic et de saisie de prescription médicale à l'hôpital

La figure 3.1 schématise ce processus de diagnostique et de prescription des soins dans un contexte hospitalier. Dans un établissement de santé, les médecins se déplacent dans l'hôpital, visitent les chambres des patients et effectuent le diagnostic, le suivi et la ré-évaluation. Suite à ce processus, le médecin saisit les prescriptions médicamenteuses en accédant à un LAP, souvent accessible uniquement par un poste fixe dans un service hospitalier connecté. Or, sur le lieu de soins ou lorsque le médecin est en déplacement, les prescripteurs n'ont pas toujours le temps ou la possibilité d'accéder au poste fixe pour saisir leur prescription. Il est donc très fréquent que la saisie ait lieu sur un support papier, écrit à la main, qui peut-être source d'erreurs iatrogènes évitables et qui rendent la traçabilité des soins plus difficile.

La saisie de prescriptions médicamenteuses n'est qu'un composant d'un processus plus large impliquant plusieurs catégories de professionnels de santé. Le circuit clinique ainsi que le circuit logistique qui compose le circuit du médicament est schématisé sur la Figure 3.2. Selon ce schéma, le médecin effectue son diagnostic et saisit une prescription. Ensuite, le médicament passe par un circuit logistique qui implique des pharmaciens d'officine ou de l'hôpital qui se chargent de son approvisionnement, de dispenser les doses à administrer et de donner les conseils de bon usage qui sont associés à cette délivrance. Enfin, le patient, ou dans certains cas d'hospitalisation des infirmiers, sont responsables de l'administration de ces doses. Suite à l'administration des médicaments, le suivi clinique permet de surveiller les effets et ré-évaluer la situation du patient.

La saisie des prescriptions informatisées est un point pivot dans l'interopérabilité des systèmes informatiques de santé et la traçabilité des soins. La saisie informatique des prescriptions, nécessitant également un logiciel paramétré par l'établissement de soins par ces différents acteurs implique une saisie contrôlée et guidée d'information afin d'effectuer certaines validations pour pouvoir alerter le personnel clinique d'éventuels risques et contre-indications.

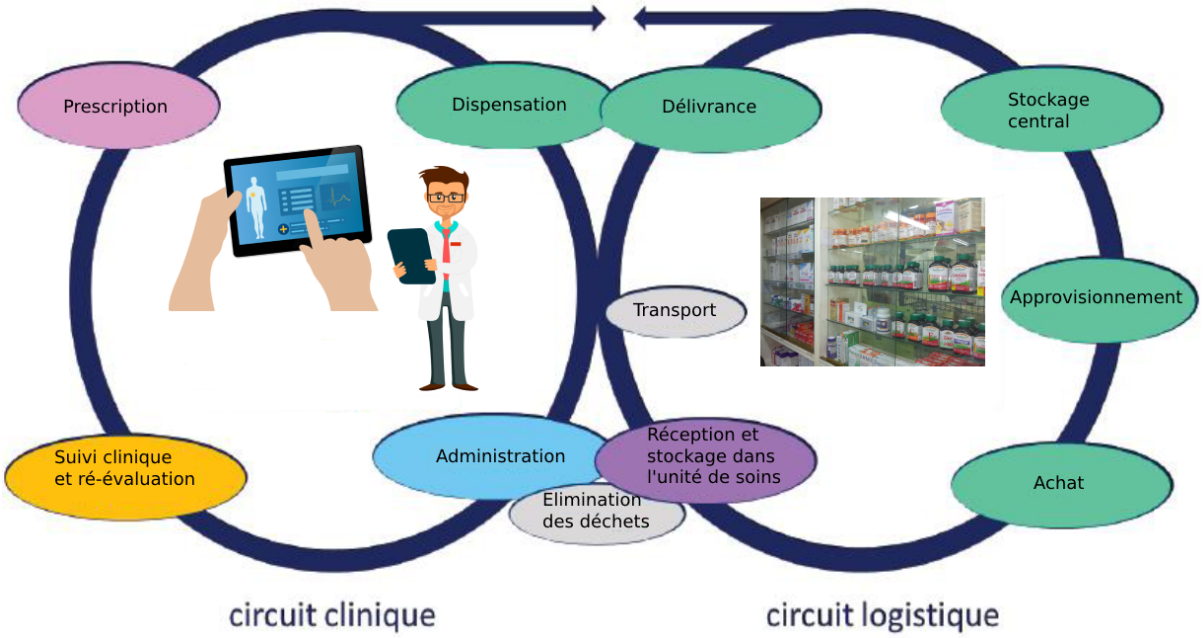


FIGURE 3.2 – Circuit du médicament adapté à partir de Gourieux (2019)

3.2 Objectif de la thèse et questions de recherche

La saisie d'ordonnance permet d'effectuer des contrôles informatiques primordiaux pour réduire les événements iatrogènes, améliorer la traçabilité et l'efficacité des soins. Cependant, ceci impose une saisie contrôlée qui ne fait pas partie de la pratique courante et représente une forme non naturelle bien que nécessaire. Cette thèse a pour objectif de libérer les prescripteurs d'un certain nombre de contraintes imposées pas la saisie informatique en permettant la saisie d'une ordonnance à l'oral. La prescription plus courante est la prescription médicamenteuse mais d'autres types de prescriptions sont possibles : des préparations médicamenteuses, perfusions, transfusions, demandes d'assistance respiratoire, analyses biologiques, examens et consultations, demandes de rééducation, et les dispositifs médicaux. Dans cette thèse, nous nous sommes concentrés sur **les prescriptions médicamenteuses** qui sont les prescriptions les plus fréquentes.

L'intérêt d'une saisie en langage naturel est multiple. La dictée vocale étant une pratique courante des professionnels de santé, une ordonnance pourrait être réalisée sur le lieu de soins. Pour cette raison nous abordons ce sujet sous forme d'une application mobile liée à un système d'information hospitalier.

Comme évoqué dans l'introduction, le bénéfice attendu serait une utilisation accrue du LAP pour d'une part, réduire les erreurs de prescription et, d'autre part, permettre une plus grande tracabilité des traitements. L'utilisation d'un terminal mobile permettrait une disponibilité accrue du système de saisie tandis que l'interface par dialogue oral éviterait les interfaces complexes actuelles et apporterait une plus grande réactivité. Hormis ces bénéfices, cette interface serait également facile d'utilisation et rapide que l'IHM clavier/souris classique.

L'un des facteurs essentiels de réussite d'un tel système est la robustesse. Le système

doit avoir un taux d'erreurs de compréhension très faible et doit fournir une résilience à celles-ci via l'intelligence humaine impliquée dans le dialogue et dont la responsabilité de prescription est engagée.

L'état de l'art montre que les systèmes les plus performants sont les systèmes à base de réseaux de neurones profonds. Cependant, bien que cela soit moins mis en avant, l'état de l'art montre également que ces modèles sont : 1) extrêmement dépendant d'un vaste ensemble de données 2) opaques et 3) imprévisibles, notamment dans un système de dialogue orienté tâche. Dans un domaine comme celui de la médecine où la nécessité de contrôler les réponses du système est primordiale, un système de bout-en-bout est inenvisageable et les systèmes en pipeline – notamment à base de règles tel que AIML– sont favorisés car interprétables et facilement liables à des ressources externes. Cependant ceux-ci sont très rigides et ne généralisent pas. Par ailleurs, dans le domaine du TALN la tâche de saisie de prescription médicamenteuse à l'oral est inexistante et, malgré nos recherches, il nous a été impossible de trouver des bases de données pour celle-ci dans aucune langue. Dans cette thèse nous avons choisi d'utiliser un système de dialogue modulaire où il est possible de faire un lien avec des connaissances médicales externes.

Dans ce chapitre, la principale question de recherche à laquelle nous nous intéressons est : **Comment concevoir un système de dialogue couplé avec un LAP dans un domaine n'ayant aucune ressource d'apprentissage?**

Pour y répondre, il est nécessaire de commencer par modéliser l'interaction conversationnelle d'une saisie. En effet, même si la pratique d'effectuer des prescriptions médicamenteuses à l'écrit est une pratique ancienne, modéliser l'interaction réelle des utilisateurs et identifier les besoins des prescripteurs d'un tel outil sont primordiaux. Or, il n'a pas d'équivalent dans la pratique médicale. Dans la section 3.3, nous allons donc répondre à la question de **comment pourrions-nous modéliser l'échange entre les différents acteurs lors de la saisie des prescriptions médicamenteuses à l'oral alors que cette pratique est inexistante?**

Nous avons détaillé les approches utilisées dans la conception des systèmes de dialogue dans le chapitre 2. En complément à cette question nous allons définir notre approche du dialogue dans le contexte d'un système de dialogue orienté but. Sur le plan dialogique, nous allons également établir les intentions du point de vue de l'utilisateur et identifier les restrictions des catégories médicamenteuses ce qui permettra de définir le domaine de notre système de dialogue.

L'un des composants le plus important d'un système de dialogue est le système de compréhension du langage. Les approches récentes abordées dans l'état de l'art nous montrent que les approches à base d'apprentissage automatique nécessitent une grande quantité de données d'apprentissage. Dans le domaine médical, notamment dans le cadre d'une pratique inexistante, l'obtention des données relève d'un défi important. Dans la section 3.4, nous allons répondre à la question suivante : **Comment utiliser l'apprentissage automatique pour obtenir des modèles d'inférence (NLU, RAP, politique de dialogue) pour un nouveau domaine qui est à priori dénué de corpus?**

La saisie des prescriptions à l'oral ne peut pas être dissociée des vérifications nécessaires pour la sécurisation des soins (p.ex., allergies du patient, médicament épuisé, etc.). Pour prendre en compte cette sécurisation des données, notre méthode prévoit une intégration avec un LAP qui devrait vérifier les informations saisies en fonction des différents profils des patients. La section 3.5 répond à cette question : **Les systèmes industriels nécessitant une maintenance et une adaptation constante devant des évolutions logicielles, légales et des industriels, comment concevoir les traitements de façon à ce que ceux reposant sur des systèmes d'apprentissages soient le moins impactés par ces changements ?**

Enfin, la section 3.6 présente les méthodes et les métriques d'évaluation utilisées pour évaluer les performances des différents composants du système. En effet, pour évaluer le système et valider l'approche il convient de se poser la question de **comment mesurer les performances de toutes les étapes du pipeline ainsi que le pipeline dans son ensemble ? Comment mesurer l'utilisabilité de ce système dans un milieu médical ?**

3.3 Définition du domaine et approche de dialogue

Nous avons vu dans l'état de l'art qu'il y a beaucoup de systèmes de dialogue émergents dans le domaine de la santé. La modélisation et la définition de ces systèmes dépendent largement de leurs objectifs et leurs publics. Une partie importante de ces systèmes sont orientés patients. Par exemple, pour un système de dialogue conçu pour apporter du soutien psychologique ou de l'aide à l'arrêt du tabac, une forme de dialogue informel « *chitchat* » pourrait être plus bénéfique pour atteindre l'objectif visé. Dans notre cas, un système de dialogue permettant de saisir des prescriptions serait orienté experts médicaux permettant de faciliter le remplissage de formulaire en ligne.

D'un point de vue global, il essentiel d'analyser les informations transmises sur les prescriptions médicamenteuses et comprendre leurs caractéristiques avant de modéliser une interaction conversationnelle potentielle avec un système. La figure 3.3 montre un exemple d'ordonnance comportant la plupart des informations nécessaires adapté à partir d'un exemple de l'Assurance Maladie (Ameli) et d'un d'ordonnance électronique complété avec une prescription manuscrite.

À part les informations administratives permettant l'identification du prescripteur et la définition du cadre légal, une ordonnance contient des « lignes de prescriptions », dont chacune est consacrée à une instruction médicamenteuse. Par exemple, sur la partie (A) de la figure 3.3 l'ordonnance comporte 7 lignes de prescriptions comportant des informations précises sur la prise du médicament, des mentions destinées aux pharmaciens et parfois les conseils associés à ces instructions. Nous pouvons remarquer sur les deux premières lignes de prescriptions que les prescripteurs peuvent prescrire un médicament en utilisant le nom commercial du médicament (nom de spécialité) ou par leur dénomination commune internationale (DCI) qui correspond au nom scientifique de la molécule active. Dans la plupart des cas, une ligne de prescription concerne la prise d'un médicament, mais dans certains cas, une ligne peut concerner la prise d'un médicament en relais avec un autre médicament

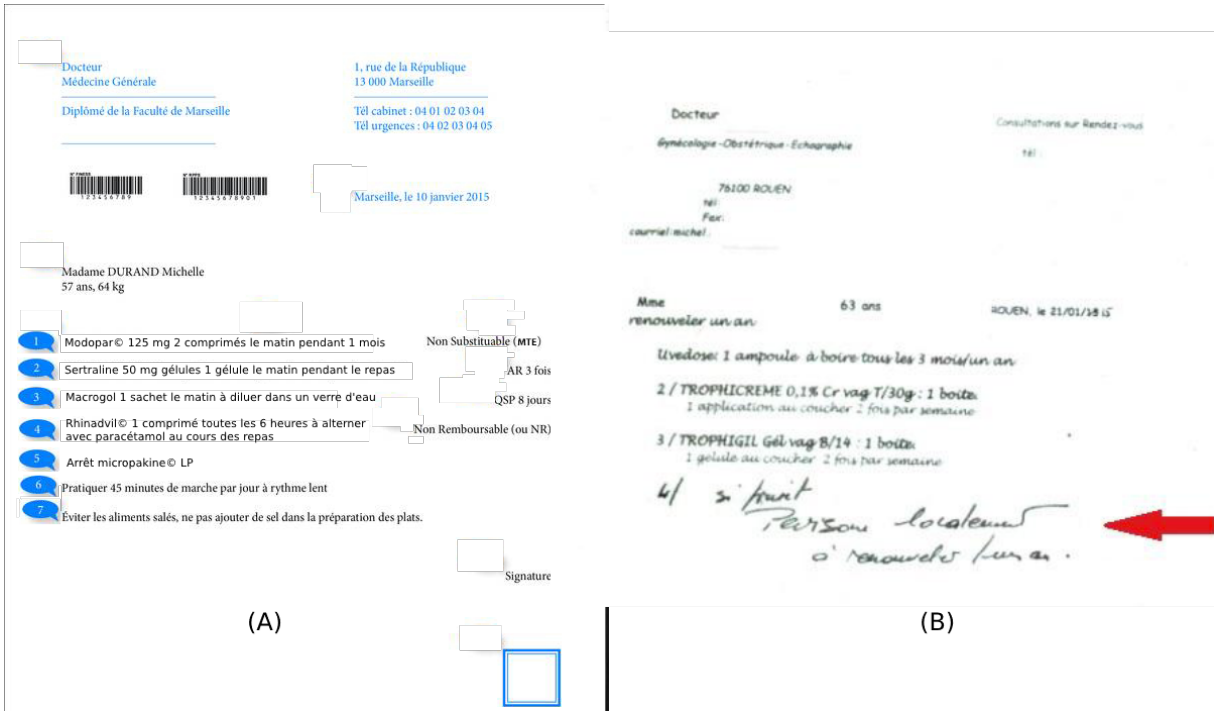


FIGURE 3.3 – (A) Éléments types d'une ordonnance inspirée d'un exemple de l'assurance maladie (Ameli) ; (B) Exemple d'ajout manuscrit qui cause une erreur de délivrance liée à la confusion dans le nom du médicament (Lachèvre, 2016).

comme sur la ligne 4. Par ailleurs, une ligne de prescription n'exprime pas forcément la prise d'un médicament, cela pourrait concerner l'arrêt ou la suspension d'un ou plusieurs médicaments comme sur la 5^e ligne de prescription. Si l'ordonnance du médicament contient des informations à corriger ou à supprimer, le prescripteur peut barrer cette ligne et/ou ajouter la mention « NPD » (ne pas délivrer) aux pharmaciens. De même, un pharmacien peut modifier une ligne de prescription s'il constate une erreur et peut modifier l'ordonnance en signant et mettant une date pour marquer des éventuelles corrections avec l'accord du médecin si nécessaire. Enfin, le prescripteur peut également instruire des conseils généraux qui n'impliquent pas la prise d'un médicament comme sur les deux dernières lignes de prescription de l'exemple (A).

L'exemple (B) de la figure 3.3 est un exemple d'ordonnance rédigée à la main. Les informations sont proches de celles écrites de manière électronique mais contiennent parfois moins de précisions. Les ordonnances manuscrites peuvent comporter également des erreurs d'écriture ou contenir des parties illisibles. Par exemple, sur la ligne de prescription (4/) de l'exemple (B), le Pevisone© crème a été ajoutée à la main par le médecin, par dessus l'ordonnance imprimée. L'étude de Lachèvre (2016) sur l'analyse des erreurs de délivrance à la pharmacie d'officine montre que cette ligne d'ordonnance a donné lieu à une erreur de délivrance à cause de l'illisibilité du nom de médicament (délivrance de Nerisone© au lieu de Pevisone©). Cette erreur de délivrance aurait pu être évitée si la totalité de l'ordonnance avait été rédigée sur un ordinateur.

3.3.1 Modélisation de la sémantique des prescriptions médicamenteuses

Les prescriptions médicamenteuses utilisent un langage très structuré et contraint tout en étant très diverses et variées. En effet, les prescriptions médicamenteuses comportent des prescriptions destinées à la fois aux patients et aux pharmaciens. Les instructions sont concises et courtes, et comportent généralement la désignation du médicament ainsi que les modalités de dosage et d'administration qu'on appelle posologie du médicament. À ces caractéristiques s'ajoutent la différence entre la communication écrite et orale. Afin de prendre en compte toutes ces caractéristiques et modéliser ce langage métier dédié à la prescription de médicaments, nous avons constitué une typologie représentant les informations essentielles figurant sur les ordonnances.

Dans le domaine médical l'utilisation des ontologies est une pratique courante qui permet l'interopérabilité sémantique (Liyanage et coll., 2015). La préoccupation majeure des ontologies du domaine médical consiste à organiser et représenter la terminologie médicale. L'une des ontologies utilisées dans le domaine du TAL, conçue pour les applications de compréhension est LinKBase® (Van Gurp et coll., 2006) comprenant un lexique de 1.5 millions de termes. Concernant les prescriptions, il existe quelques ontologies sur la sémantique d'une prescription telles que PDRO [1] (The Prescription of Drugs Ontology) qui représente les informations liées au médicament et la posologie de façon structurée (Aronson, 2001; Ethier et coll., 2016). Cependant, ces ontologies ne sont pas adaptées pour représenter les prescriptions médicamenteuses formulées en langage naturel. La plupart des ontologies sont réalisées en vue de représenter les données pour les systèmes informatique de santé, c'est-à-dire une fois qu'elles sont formalisées.

Pour représenter la sémantique des prescriptions, dans un processus itératif, nous avons étudié les différentes ontologies proposées dans le domaine, des exemples de prescriptions issus d'un livre académique qui sera détaillé dans la section 3.4, les connaissances issues d'un thésaurus médical et les LAP avec un spécialiste du domaine. La démarche que nous avons suivie consiste à :

1. Analyser des bases de données d'un LAP en lien avec un thésaurus médical
2. Extraire et classer des exemples de prescriptions en catégories
3. Proposer une description permettant de décrire de manière exhaustive les exemples
4. Confronter cette description avec un expert du domaine, si non valide retourner en 1.

Cette approche itérative nous a permis de créer une typologie détaillée définissant la sémantique des prescriptions médicamenteuses qui est apte à représenter les informations exprimées en langage naturel. Schématisés sur la Figure 3.4, les cadres sémantiques sont regroupés selon ces grandes catégories :

— **Désignation du médicament** : Le nom de spécialité ou les substances actives du médicament

1. http ://purl.obolibrary.org/obo/PDRO_0000024

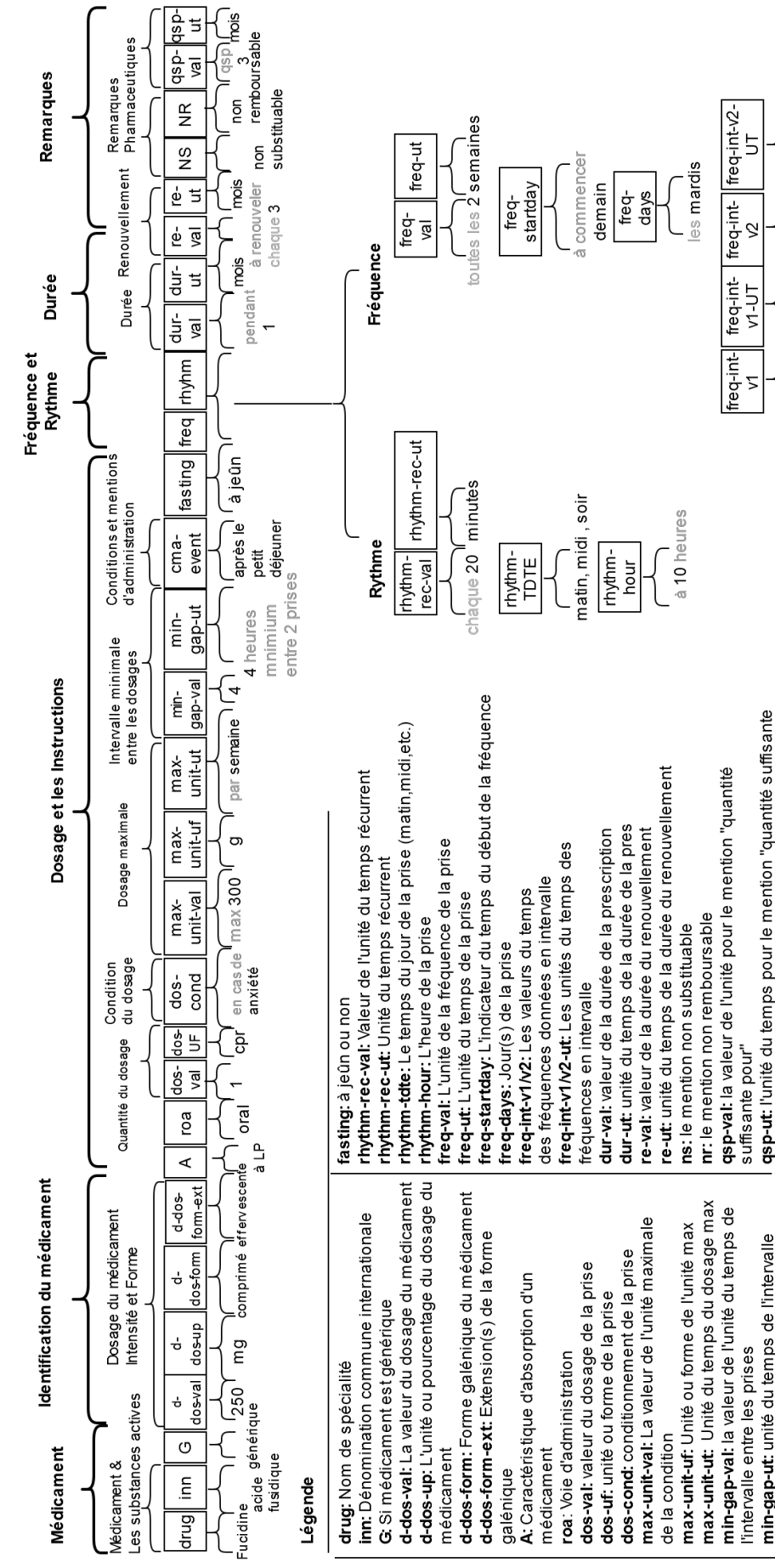


FIGURE 3.4 – Typologie du domaine des prescriptions médicamenteuses avec des exemples de valeurs

— **Identification du médicament** : Les informations liées au dosage, à l'intensité et la forme galénique du médicament.

— **Le dosage et les instructions** : Toutes les informations qui concernent le dosage destiné pour le patient. Cela concerne la quantité du dosage, s'il s'agit d'une prise conditionnelle (ex : en cas de douleurs), les conditions de la prise, l'espacement minimal entre deux prises, le dosage maximal à ne pas dépasser en cas de posologies adaptables en fonction des symptômes des patients et les conditions et mentions d'administration qui peuvent être plus larges (ex. prendre le cachet en mangeant le repas du soir).

— **Rythme de la posologie** : Les informations liées aux heures de la prise dans une journée. Par exemple, la prise d'un médicament à heure fixe ou dans un temps moins précis comme matin, midi et soir ou encore une expression exprimée dans un intervalle temps journalier (ex : toutes les 2 heures).

— **Fréquence de la posologie** : L'expression de la fréquence de la prise. La fréquence la plus typique est tous les jours, mais il est possible de préciser des fréquences et intervalles différents. Par exemple, un prescripteur peut vouloir prescrire une injection une fois par mois ou laisser une fenêtre thérapeutique de quelques jours dans la fréquence journalière (ex : prise d'un médicament dans les jours de la semaine uniquement)

— **Durée** : Les informations concernant la durée totale et les informations liées au renouvellement de l'ordonnance.

— **Remarques** : Les remarques pharmaceutiques destinées aux pharmaciens comme *non substituable* ou *non remboursable*, etc.

Les attributs sémantiques de chaque catégorie décrits ci-dessus sont détaillés sur la Figure 3.4. Ces attributs sémantiques sont décortiqués pour la plupart en sous-catégories fines distinguant la valeur numérique et la désignation *(1=dos_val, comprimé=dos_uf)*, afin de permettre au système de dialogue d'adapter la politique en fonction des erreurs de compréhension et de la reconnaissance de la parole.

Un logiciel d'aide à la prescription est soumis à un certain nombre de contraintes et réglementations de la Haute Autorité de Santé (HAS). Dans l'approche de remplissage d'attributs, les valeurs récupérées par les modèles de compréhension peuvent être quasiment illimitées. Afin de restreindre les valeurs des différents cadres ainsi que pour permettre une meilleure catégorisation, nous avons défini des attributs-valeurs normalisés. De manière générale, nous pouvons classifier les valeurs portées par les attributs définis sur la Figure 3.4 en trois catégories : en premier, les attributs qui portent des valeurs numériques. Par exemple, l'attribut « d_dos_val » représente la valeur du dosage du médicament et doit représenter des valeurs numériques. Le deuxième type de valeur concerne les valeurs qui font partie d'une liste finie ; par exemple, les unités de prescription exprimées en unité unités de mesure (mg, kg, etc.). Enfin, nous avons la catégorie d'information la plus difficile à représenter, celle qui englobe des informations qui peuvent être très variées comme les conditions de la prescription.

Parmi les attributs, les principaux sont sans doute la désignation d'une ou plusieurs substances actives à effet thérapeutique. Cette désignation se fait en général avec le nom commercial, c'est-à-dire le nom de spécialité du médicament ou sa DCI. Cependant, il est également possible de se référer à un médicament en indiquant le nom du générique, le nom d'un produit vendu en vente libre dans une pharmacie, le nom collectif d'une substance à effet thérapeutique ou de la substance biologique (Uzuner et coll., 2010b).

De plus, les valeurs de ces attributs sont à la fois limitées parce qu'il s'agit des médicaments qui existent mais qui évoluent avec le temps. Les nouveaux principes actifs et les médicaments qui entrent dans le marché font partie de ces valeurs et représentent une difficulté pour les systèmes d'inférence (Liu et coll., 2015). Les approches bien connues dans le domaine de la reconnaissance d'entités nommées dans le domaine biomédical sont MetaMAP (Aronson, 2001) ou RxNorm (Nelson et coll., 2011) qui est spécialisé sur les médicaments. Ces systèmes proposent des noms normalisés, des identifiants uniques, et plusieurs niveaux de description et de relations.

Il est primordial pour les logiciels d'aide à la décision et les LAP d'utiliser une ontologie permettant l'interopérabilité sémantique. Pour prendre en compte toutes les particularités du langage naturel et les énoncés oraux pour le français, nous nous sommes basés sur un thésaurus médicale autorisé par la HAS : Thesorimed©. Similaire aux ontologies et autres thésaurus (tels que UMLS (Bodenreider, 2004)), cette base de données médicamenteuses propose des informations normalisées et permet d'identifier des médicaments avec les codes d'identification utilisés en France permettant une interopérabilité sémantique avec d'autres systèmes autorisés par la HAS. Concernant les valeurs normalisées des différents attributs sémantiques, nous nous sommes basés sur les valeurs finies comme les unités de mesure ou les voies d'administration, etc. proposées par Thésorimed et en consultant un expert dans le domaine médical.

3.3.2 Conception du flux dialogique

La validation et l'intégration d'une prescription médicamenteuse dans un plan de soins nécessiterait plusieurs niveaux d'interaction. La figure 3.5 schématise de façon globale les différents niveaux d'interaction que nous avons conçus avec un exemple de dialogue.

Le point de départ du dialogue est la transcription de l'énoncé oral correspondant à une prescription ou à des éléments d'informations d'une prescription médicamenteuse. Une fois que le prescripteur initie le dialogue, l'étape **1** (Figure 3.5) du système de dialogue consiste à extraire la sémantique de l'énoncé par le remplissage d'attributs (*slot-filling*). Pour chaque attribut le système extrait une étiquette et une valeur. Par exemple, la forme galénique de la posologie comporte l'étiquette et la valeur suivante : *dos_uf = (tablet,"comprimé")*.

Les énoncés des prescripteurs étant concis et précis, un complément et une vérification des informations liées au médicament sont nécessaires. L'attribut le plus important d'une prescription est le médicament : son nom commercial (Xanax©) ou sa dénomination commune internationale (ALPRAZOLAM). Comme un médicament peut avoir plusieurs entrées

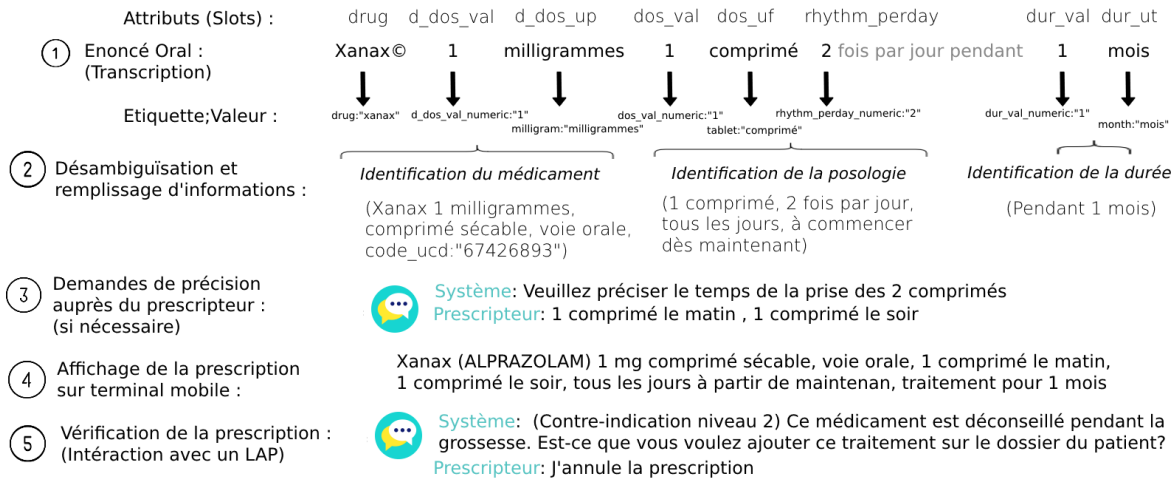


FIGURE 3.5 – Schéma des étapes globales du système de dialogue

dans une base de données médicamenteuses (par exemple, un même médicament peut exister sous forme de gélule ou de granulé, avec différents dosages, etc.) il est essentiel d'accumuler toutes les informations sur le médicament afin de désambiguïser et permettre son association avec une entrée unique. L'étape **2** de la figure 3.5 illustre cette étape qui permet de compléter les informations de manière formelle (Xanax©1 mg, comprimé sécable, ...). S'il y a plusieurs médicaments correspondant aux attributs de l'état actuel du dialogue, le système renvoie une liste de médicaments proposés au prescripteur.

Une fois que le médicament et les informations le concernant sont identifiés, l'étape **3** consiste à compléter les informations nécessaires pour la prescription : la posologie et la durée de la prescription. Selon la définition de notre domaine, la plupart des informations sur les prescriptions sont des informations dites non obligatoires. Par exemple, le fait qu'un médicament soit remboursable ou non est un bon exemple d'une information porteuse de sens pour une prescription mais qui comporte un caractère optionnel. Dans cette étape, le module de gestion de dialogue identifie ces informations obligatoires et les associe à l'acte de dialogue adéquat qui permettra de les collecter.

Une fois que les informations liées à la posologie, la durée et le médicament sont complètes, le système effectue des vérifications liées à la cohérence de la prescription. Dans cet exemple, *2 fois par jour* n'est pas assez précis, mais constitue un point d'ancrage pour le dialogue afin de confirmer s'il s'agit d'une prise le matin et le soir. Après les précisions de l'utilisateur, si nécessaire, l'étape **4** de la figure 3.5 entre en jeu : la prescription complète sera répétée à l'utilisateur pour une validation explicite de celle-ci. Les informations sont affichées également de façon explicite sur le terminal mobile pour que le prescripteur puisse les vérifier une dernière fois.

L'une des fonctionnalités majeures d'un LAP est sa fonctionnalité de vérification des interactions médicamenteuses et d'autres informations liées au médicament et selon le profil du patient. Dans cette dernière étape **5**, après l'ajout de la prescription, un LAP est capable de renvoyer une alerte auprès des prescripteurs précisant le motif de l'alerte. Cette étape constitue la dernière étape principale de notre système de dialogue. Si le LAP ne trouve pas

de contre-indications, la prescription est ajoutée sur le dossier du patient. Le prescripteur peut également valider la prescription pour l'ajouter sur le dossier après avoir pris connaissance des éventuelles contre-indications en validant le retour du LAP.

3.3.3 Définition des actes de dialogue

Comme décrit dans la sous-section précédente, la prescription médicamenteuse intégrée à un LAP nécessite plusieurs niveaux de contrôles et de vérification. Afin de concrétiser cette distinction entre un système de dictée vocale et un système de dialogue, nous avons établi les intentions potentielles du système et de l'utilisateur. En effet, avec l'approche de *slot-filling*, le système extrait les attributs sémantiques mais détermine également l'intention de l'utilisateur pour que le système puisse donner une réponse dans l'état actuel du dialogue. Ces intentions correspondent aux représentations informatiques des actes de dialogues envisagés.

Comme nous l'avons vu dans l'état de l'art, un système dialogique utilise la notion de *tours de parole*. Dans notre système orienté but, un tour de parole est composé d'un tour d'utilisateur et d'un tour de système. C'est-à-dire que tous les énoncés utilisateurs doivent être suivi d'une réponse (orale, visuelle ou sous forme d'action externe). Le tableau 3.1 résume les actes de paroles dans le cadre du dialogue que nous avons défini.

Acte de paroles	S	U	Description
inform(task=pres-drug, a=x,b=y,...)		✓	commencer le dialogue par une prescription médicamenteuse avec une ou plusieurs informations liée à la prescription
request(a=restart)		✓	l'utilisateur informe explicitement le souhait de recommencer la prescription en cours
inform(a=drug_not_found)	✓		lorsque le médicament est introuvable dans la BD, le système informe l'utilisateur
inform(a=drug,b=dos-val,...)		✓	l'utilisateur donne l'information a=x, b=y, ...
ask(a=dos-val || a=dos-uf,...)	✓		le système demande à l'utilisateur de compléter une information manquante
affirm()		✓	l'utilisateur valide la prescription
negate()		✓	l'utilisateur refuse la prescription
repeat()	✓		le système demande de répéter la dernière acte de dialogue
reqalts(a=drug,b=d-dos-val,...)	✓		le système propose une liste des alternatives étant donné a=X ,b=Y, ...
request(a=comment)	✓		le système demande à l'utilisateur d'ajouter un commentaire ou une observation sur la prescription en cours
inform(a=comment,b=x)		✓	l'utilisateur ajoute un commentaire libre sur la prescription en cours
warn(a=x,drug=y,inn=z,...)	✓		étant donné les informations de la prescriptions, le système alerte l'utilisateur sur des EIM potentiels
negate(a=x)		✓	l'utilisateur nie l'information(a=x)
negate(a=x,b=y)		✓	l'utilisateur nie l'information (a=x) et corrige l'information comme (b=y)

TABLE 3.1 – Résumé de la définition des actes de dialogue définis dans le système (*S=Système, U=Utilisateur*).

Le tableau 3.1 présente les actes de dialogues avec les attributs et des exemples de valeurs, leur catégorie ainsi que leur description. L'un des avantages d'un smartphone est la possibilité d'afficher des informations sur l'écran. Cet avantage a été pris en considération

pour représenter aussi des actes physiques correspondant aux actes de parole lorsque cela est pertinent. Par exemple, l'affichage de la liste des médicaments possibles d'une prescription est plus facile à visualiser sur un terminal mobile qu'à exprimer par voix de synthèse et le choix du prescripteur peut plus facilement être exprimé par un clic que par l'énonciation.

Comme catégorie de prescription médicamenteuse, nous nous sommes concentrés sur l'intention de haut niveau : prescriptions médicamenteuses. Les actes de paroles définis ci-dessus concernent la prescription, la correction, la validation et l'annulation des lignes de prescriptions. Ces intentions de haut niveau sont également représentées dans l'ontologie PDRO avec une représentation légèrement différente. Dans cette ontologie, la prescription médicamenteuse est définie selon ces 4 catégories : l'initiation, la poursuite, la modification ou l'arrêt de l'administration du médicament (Ethier et coll., 2016). Contrairement à l'ontologie PDRO, notre définition sémantique ne distingue pas les attributs selon ces caractéristiques mais les représente comme un problème de classification d'intention où l'utilisateur construit l'ordonnance au fur et à mesure, guidé par le système. Nous avons cependant cerné le problème en limitant les prescriptions à un seul médicament par ligne de prescription. De même, nous avons limité la référenciation à d'autres lignes de prescription ou d'autres traitements en cours du patient afin de se concentrer sur la prescription d'un médicament à l'oral.

3.3.4 Approche itérative de la modélisation du dialogue

La modélisation conversationnelle dans un domaine spécialisé nécessite la consultation des experts du domaine, notamment dans le cadre d'une pratique inexistante. De ce fait, nous avons d'abord étudié le domaine et proposé une typologie ainsi que les intentions utilisateurs et avons validé ces choix avec un expert du milieu médical.

Afin de valider l'approche et la pertinence des actes de dialogue et les stratégies dialogiques des utilisateurs, nous avons procédé de façon itérative. Suite à la phase de la modélisation théorique et de la validation auprès d'un expert, nous avons développé un prototype initial pour A) collecter des données et B) observer les stratégies employées par les experts du domaine et les utilisateurs naïfs dans un contexte dialogique.

En effet, face aux erreurs de reconnaissance de la parole ou de compréhension, il est important d'observer la réaction des experts du domaine. Par exemple, face à une erreur, quelle serait la stratégie pour la corriger, est-ce que le prescripteur préférera tout recommencer ou modifier seulement un élément ? Est-ce qu'il cherchera à cliquer sur l'erreur pour la corriger ou essaiera de la modifier à l'oral ? Pour obtenir des réponses à ces questions ainsi qu'adapter les actes dialogiques, nous avons développé un premier prototype en 2018 pour valider l'approche de la compréhension du langage (sans collecte dialogique). La méthodologie sur la compréhension automatique de la sémantique des prescriptions sera détaillée dans la section 3.4. Suite à la validation de l'approche de compréhension, nous avons mis en place un prototype initial permettant la collecte et une petite évaluation validant l'approche. Suite à cette validation, nous avons adapté le prototype, amélioré le système de compréhension

afin d'établir un prototype pour collecter du corpus oral à plus grande échelle. Le tableau 3.2 résume les étapes de cette approche itérative.

Date	Phase	Notes
Fin 2018	Premier prototype sur la compréhension des prescriptions médicamenteuses	
Fin 2019	Premier prototype de dialogue	Expérience initiale avec 2 utilisateurs naifs 2 experts médicaux
Décembre 2020	Validation du protocole de la collecte et la préparation du prototype pour une collecte à plus grande échelle	Validation du protocole avec 4 experts du domaine
Décembre 2020-Juillet 2021	Collecte d'un corpus oral dans un contexte dialogique	
Juin 2021	Premier prototype intégré avec un LAP	

TABLE 3.2 – Calendrier des prototypes mis en place dans le cadre de l'approche itérative

Le premier prototype figurant sur le tableau 3.2 qui a commencé à la fin 2018 faisait suite au processus itératif décrit dans la section 3.3.1. Ce premier prototype nous a permis de tester la méthode de remplissage d'attributs, d'intégrer la reconnaissance automatique de la parole et la synthèse vocale et de valider la typologie avec les experts du domaine sur un terminal mobile. Nous avons également comparé les performances des différents modèles de compréhension sur un ensemble de données qui sera décrit dans la section 3.4.

Ensuite, nous avons modélisé l'interaction conversationnelle du domaine des prescriptions médicamenteuses. Nous avons appliqué la même méthodologie itérative pour cette phase de modélisation conversationnelle. Au fur et à mesure, nous avons adapté le prototype mobile pour le faire évoluer dans le sens d'un système de dialogue. Ce prototype a donné lieu a une évaluation intermédiaire à petite échelle afin d'analyser les choix méthodologiques avec deux utilisateurs naïfs et deux utilisateurs experts fin 2019.

Après que le système de dialogue a été revu plusieurs fois suite aux avis des experts, afin de permettre une évaluation et en même temps de constituer un corpus oral à la fois avec des utilisateurs naïfs et les experts du domaine médical, nous avons préparé un protocole de collecte qui était soumis à la CNIL. Suite à la validation du protocole avec 4 experts médicaux, nous avons commencé une campagne d'évaluation en décembre 2020 qui a continué jusqu'au mois de juillet 2021. Au début, la collecte était prévue avec la présence d'un présentateur en présentiel avec du matériel fourni aux participants pendant la collecte. Cependant, avec le contexte sanitaire et l'épidémie de Covid-19, nous avons adapté le protocole pour permettre une collecte qui a pu être réalisée complètement à distance.

A la fin, nous avons également proposé un prototype plus détaillé connecté avec les dossiers électroniques des patients fictifs et capable d'effectuer les vérifications nécessaires à l'aide d'un logiciel de LAP permettant d'alerter les prescripteurs des éventuelles contre-indications au mois de Juin 2021.

3.4 Approche de la compréhension du langage dans le domaine des prescriptions médicamenteuses

Dans la section 3.3 nous avons décrit notre approche de la modélisation de la sémantique du domaine des prescriptions médicamenteuses ainsi que notre approche de la modélisation dialogique. L'un des composants majeurs d'un système de dialogue est le système de compréhension car afin d'aboutir vers un dialogue, il faut d'abord que le système soit capable d'extraire la sémantique à partir des énoncés. La compréhension automatique de la sémantique consiste à extraire des concepts à partir des cadres sémantiques (Fillmore et coll., 1976) dont les cadres (aussi appelés *slots* représentant des bouts d'informations pertinentes. Les cadres représentent ces bouts d'informations sous des catégories restreintes par une définition fine d'un espace sémantique (Tur et De Mori, 2011). Pour ces bouts d'éléments qui sont pertinents à l'énoncé, on distingue la catégorie sémantique (slot-label) ou son étiquette et sa valeur normalisée (slot-valeur) ou son étiquette-valeur. Dans un contexte dialogique, notamment dans le cadre d'un système de dialogue orienté but, nous ne pouvons pas dissocier la sémantique extraite de l'intention du locuteur. L'intention du locuteur pourrait être définie comme l'action (implicite ou explicite) que l'utilisateur veut effectuer. Cette notion approche celle de l'acte de langage vu dans l'état de l'art et est étroitement liée à la définition du domaine décrite dans la section 3.3. Nous abordons la compréhension du langage avec cette approche de remplissage d'attributs ou *slot-filling* schématisée sur la Figure 3.6.

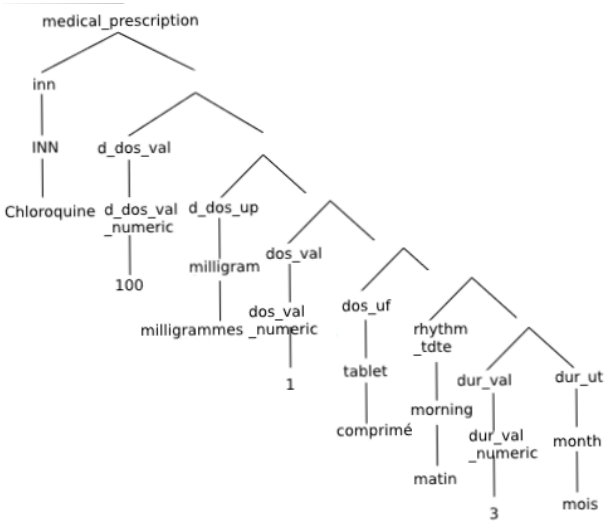


FIGURE 3.6 – Représentation hiérarchique de concepts du domaine des prescriptions médicamenteuses

La prescription « *Chloroquine 100 milligrammes 1 comprimé le matin pendant 3 mois* » est représentée sous forme hiérarchique sur la Figure 3.6. La particularité des prescriptions médicamenteuses est leur nature condensée : des énoncés courts comportant beaucoup d'information sémantique. En effet, il faut qu'une prescription soit concise et compréhensible par les pharmaciens, patients et d'autres médecins. Nous pouvons remarquer que

même dans cet exemple court, le nombre de cadres extraits est élevé par rapport au nombre de mots (8/10).

Les approches récentes utilisées dans le domaine de la compréhension du langage s'appuient sur l'apprentissage de grandes quantités de données. Or, dans une nouvelle tâche comme celle-ci, il n'existe pas de données prêtes à entraîner. Les défis dans le domaine biomédical (pour le français ou l'anglais) abordés dans l'état de l'art se concentrent sur l'extraction de la sémantique autour d'un médicament à partir des comptes rendus médicaux. En revanche, dans le cadre d'une interaction avec un agent conversationnel, le système a besoin de reconnaître non seulement la sémantique des prescriptions complètes mais également ? des bouts d'informations permettant de compléter et corriger des prescriptions ou d'extraire de la sémantique aux réponses données par les utilisateurs à des éventuelles questions du système.

Pour obtenir des données dans ce nouveau domaine dénué de corpus, nous avons tout d'abord analysé les livres pédagogiques afin de trouver des livres contenant des exemples de prescriptions médicamenteuses. Les prescriptions que nous avons extraites automatiquement à partir des ressources pédagogiques nous ont servi à étudier le domaine et à constituer un premier ensemble de données pour entraîner des modèles de compréhension. La méthodologie de cette approche pour constituer une source de données initiales sera détaillée dans la partie 3.4.1.

En revanche, l'extraction de quelques centaines de prescriptions n'est pas suffisante pour entraîner des modèles d'inférence dans un nouveau domaine. De plus, les prescriptions extraites contiennent beaucoup d'exemples de prescriptions typiques avec quelques exemples de prescriptions plus complexes. En effet, les prescriptions simples couvrent la grande majorité des cas de prescriptions statistiquement mais un LAP devrait être toujours capable de représenter les prescriptions différentes de la prescription des médicaments habituels. De la même manière, la sémantique représentée par ces exemples tendent plus ? vers certains attributs que d'autres. Afin d'équilibrer cet ensemble de données et obtenir une quantité suffisante de données nous avons généré des données artificielles prenant en compte la représentation des attributs. Les spécificités et la méthodologie de cette approche de génération sont détaillées dans la section 3.4.2.

3.4.1 Sources de données de prescriptions médicamenteuses

Les avancées récentes décrites dans l'état de l'art tirent l'avantage de la disponibilité d'un ensemble de données annotées et la puissance de calcul permettant aux modèles d'inférence d'avancer dans tous les domaines du TAL. En revanche, dans le domaine médical, un domaine qui implique la vie privée des personnes, la disponibilité des entrepôts de données (dataset) accessibles librement sont rares. Pour les systèmes à base d'approches statistiques, les données occupent une place primordiale car elles constituent des exemples fournis au système pour l'apprentissage d'une tâche spécifique. Pour l'anglais, il existe des datasets sur la santé comprenant des informations sur les prescriptions comme nous l'avons déjà abordé

dans l'état de l'art. Cependant, ces corpus sont en anglais et ne comportent pas de prescriptions formulées en langage naturel. Pour l'entraînement d'un système de dialogue oral, l'entraînement nécessite des données annotées en langage naturel et de préférence à l'oral. À notre connaissance, il n'existe pas de corpus de prescriptions médicamenteuses en langage naturel disponible qui pourrait être utilisé dans l'entraînement des systèmes.

La collecte de données initiales avait un double objectif : cela permettrait une modélisation de la sémantique et les exemples annotés pourront être utilisés comme un corpus d'apprentissage. De plus, cet échantillon d'exemples de prescriptions pourrait servir pour catégoriser les informations pertinentes dans des cadres sémantiques.

Notre stratégie de départ était de nous orienter vers des livres destinés aux étudiants en médecine qui pourraient contenir des prescriptions. Certains guides sur le diagnostic, des thérapeutiques et des cas cliniques comportent des exemples de prescriptions mais sont détaillés d'une façon pédagogique. Le livre que nous avons choisi pour l'extraction des données initiales est un livre destiné aux étudiants en médecine comprenant des types de prescriptions selon les spécialités : « Mon Guide des Premières Ordonnances » (Lariven, 2008).

Bien qu'on ne dispose pas de corpus d'apprentissage en langage oral comportant des formulations naturelles des prescriptions médicamenteuses de différents types, nous avons obtenu des exemples variés de différentes spécialités de médecine. La plupart des prescriptions étaient cadrées en dehors du paragraphe de texte ce qui a facilité l'extraction avec les expressions régulières. L'approche que nous avons adoptée pour extraire ce premier échantillon de prescriptions comportait les six étapes suivantes :

1. Extraction manuelle du contenu des différents chapitres : Cette étape a consisté à étudier les chapitres afin d'exclure ceux comportant des prescriptions très spécialisées ou ceux comportant beaucoup de notes pédagogiques. Par exemple, nous avons exclu le chapitre nutrition car les prescriptions étaient majoritairement non médicamenteuses.

2. Extraction automatique des prescriptions avec les expressions régulières : Dans Lariven (2008) la plupart des ordonnances étaient séparées du texte dans des rubriques dédiées aux prescriptions, ce qui a facilité l'extraction automatique par l'identification des marqueurs du début des prescriptions comme « — Paracétamol... »

3. Nettoyage et normalisation par des expressions régulières : Cette étape de l'extraction a permis de nettoyer au maximum les notes pédagogiques destinées aux étudiants. En effet, vu qu'il s'agit d'un livre destiné à des fins pédagogiques, beaucoup de conseils ou de choix sont présentés comme à adapter en fonction du profil des patients. Dans cette étape nous supprimons les informations données entre les parenthèses ou les détails donnés sur plusieurs phrases. À part le nettoyage, nous appliquons également des motifs de recherche et de remplacement pour les acronymes et écritures raccourcies comme *(cpr, comprimé)*. Les posologies précisées sous forme de *x mg/kg* sont remplacées par des valeurs inventées afin d'obtenir un résultat qui ressemble le plus à une prescriptions finale.

(1) DEROXAT [paroxétine] cp 20 mg : posologie initiale de 20 mg par jour. Après un délai de 3 semaines et en fonction de l'efficacité, cette posologie peut être augmentée par paliers de 10 mg jusqu'à un maximum de 50 mg par jour.

(2) DEROXAT cp 20 mg : 20 mg par jour

FIGURE 3.7 – Exemple de prescription extraite à partir du Lariven (2008) et le résultat de l'extraction et nettoyage automatique

4. Mélange automatique des prescriptions et extraction sous un format prêt à annoter : Vu que chaque chapitre correspond à une spécialité en médecine, les prescriptions et les médicaments sont plus proches les uns des autres au sein d'un chapitre. En dernière étape, pour ne pas induire un biais sur le corpus, nous mélangeons les prescriptions avant l'extraction finale sous un format standard prêt à annoter.

5. Annotation manuelle du corpus : En utilisant un outil web pour faciliter le processus, nous avons procédé à l'annotation de la sémantique (étiquettes et étiquettes-valeurs) des prescriptions extraites.

6. Revoir le corpus avec un expert du domaine, pour les annotations non valides retourner en étape 5.

Dans ce livre, les prescriptions sont souvent accompagnées des notes et des alternatives à adapter en fonction des profils des patients. Les prescriptions peuvent faire partie d'un plan de soin constitué de différents niveaux de traitements. Par exemple, une ordonnance peut comporter la prise d'un médicament en première intention et comporter une deuxième ligne de prescription à prendre en cas de crise ou d'insuffisance de la première intention. Dans notre approche dialogique, vu qu'on restreint la prescription d'un médicament à la fois, le nettoyage permet de supprimer les conditions complexes de substitution, ré-évaluation et de suivi clinique et le mélange automatique permet de rendre les prescriptions indépendants les unes des autres.

La Figure 3.7 illustre un exemple de prescription extraite automatiquement et le résultat de la prescription suite au nettoyage automatique avec les expressions régulières. Dans cet exemple de la Figure 3.7, les notes pédagogiques destinées aux étudiants sont exclues par les expressions régulières ce qui donne le résultats des prescriptions plus courtes mais plus proches de celles qui sont rédigées sur une ordonnance. Dans cette étape de nettoyage, nous excluons également les noms des DCI qui sont systématiquement donnés entre les crochets carrés qui constituent la plupart des prescriptions. Nous séparons également l'indication de plusieurs médicaments sur une même ligne lorsqu'ils sont séparés avec les conjonctions de coordination (clonazépam gouttes ou amitriptyline 2%...)

3.4.2 Génération automatique des prescriptions médicamenteuses

Les informations figurant sur une prescription sont vastes. Cependant, la plupart des prescriptions, au moins celles extraites du Lariven (2008), comportent un sous-ensemble de ces informations. Malgré tout, les données d'entraînement devront être équilibrées pour couvrir toutes les situations, même les plus rares. Avant de constituer le corpus et d'effectuer la répartition pour l'entraînement, il a fallu extraire les statistiques des attributs. Le tableau 3.3 illustre les cadres plus fréquents et d'autres qui le sont moins.

Attributs fréquents	**Attributs rares**
drug : 409	inn : 29
rhythm-perday : 105	rhythm-rec-val : 2
d-dos-form : 68	d-dos-form-ext :3
dur-val : 108	re-val : 0

TABLE 3.3 – Exemples d'attributs fréquents et d'attributs rares

La fréquence de certains cadres est beaucoup plus importante que d'autres. Comme décrit dans la partie précédente, les DCI (inn) systématiques marqués entre parenthèses après chaque spécialité ont été exclus pendant le pré-traitement. En effet, il aurait été possible de garder les DCI pour avoir une répartition égale entre les attributs drug et inn mais cela aurait dupliqué le nombre de prescriptions. Nous avons retenu seulement les DCI lorsque la prescription était faite initialement en DCI.

Ce phénomène de répartition non-uniforme des classes est un phénomène courant dans le domaine de l'apprentissage machine. Pour résoudre ce problème, il existe plusieurs solutions comme la pondération des exemples durant l'apprentissage ou la génération de données via des réseaux antagonistes génératifs (Fedus et coll., 2018; Golovneva et Peris, 2020) (GAN en Anglais). Cependant, le corpus acquis est trop petit pour appliquer ces méthodes. De plus, la pondération ne résout pas le problème de couverture des exemples et l'augmentation stochastique des données nécessite un ensemble initial de données trop important. Par conséquent, pour l'augmentation des données et afin de résoudre le déséquilibre des classes, nous avons mis en place une technique de génération simple. Cette technique consiste à rédiger une grammaire hors contexte à base de traits pour représenter les éléments composant une prescription. Malgré le manque de structures variées, cela nous a permis d'obtenir une quantité de données riche en information sémantique pour augmenter les données de prescription.

Pour adapter au contexte dialogique, nous avons conçu les règles sous les regroupements sémantiques décrits précédemment dans la section 3.3.1. De cette manière, des bouts d'informations figurant sur les prescriptions peuvent être générés indépendamment les uns des autres. Sur la définition du domaine figurant sur la Figure 3.4, ces regroupements sémantiques sont : médicament, identification du médicament, dosage et instructions, fréquence et rythme, durée et remarques pharmaceutiques. Dans l'idéal, un système de dialogue serait capable de reconnaître et répondre aux énoncés qui ne sont pas des prescriptions complètes.

Pour permettre cela et pouvoir générer des exemples pertinents représentant les sous dialogues, le langage engendré par notre grammaire à traits représente les regroupements sémantiques. Ces regroupements sémantiques constituent des prescriptions complètes lorsqu'ils sont concaténés. La Figure 3.8 montre un extrait de règles concernant le regroupement sémantique sur le médicament et son identification.

```
# Regle (1)
S[intent=Drug_Prescription, DRUG=[Drug=?d,INN=?i,G=?g],
                DrugID=[ddosval=?dd,ddosup=?du,ddosf=?f,ddosfext=?fe,A=?a]] ->          DAct_Drug[Drug=?d,G=?g,INN=?i] \
                                                                                    DAct_DrugIdentification[ddosval=?dd, \
                                                                                    ddosup=?du,ddosf=?f,ddosfext=?fe,A=?a]

# Regle (2): Doliprane
DActDrug[Drug=?d,G=?g,INN=?i] -> SPEC[Drug=?d]
# Regle (3): Doliprane Paracetamol
DActDrug[Drug=?d,G=?g,INN=?i] -> SPEC[Drug=?d] INN[INN=?i]

# Regle (4) Doliprane 20 mg
DActDrugIdentification[ddosval=?dd,\
ddosup=?du,ddosf=?f,ddosfext=?fe,A=?a] -> DOSAGE[ddosval=?dd,ddosup=?du]
# Regle (5) Doliprane 20 mg comprime pellicule a LP
DActDrugIdentification[ddosval=?dd,\
ddosup=?du,ddosf=?f,ddosfext=?fe,A=?a] -> DOSAGE[ddosval=?dd,ddosup=?du] \
                                          GALENICFORM[t=complex,ddosf=?f,ddosfext=?fe] \
                                          ABSORPTION[A=?a]
# Regle (6)[int,dec] mg
DOSAGE[ddosval=?dd,ddosup=?du] -> NUMERICVALUE[t=decint, ddosval=?dd] \
                                  UNIT[t=?s,ddosup=unit]
# Regle (7) [int,dec] %
DOSAGE[ddosval=?dd,ddosup=?du] -> NUMERICVALUE[t=decint, ddosval=?dd] \
                                  PERCENT[ddosup=?du]
# Terminaux
SPEC[Drug=drug] -> '<SPEC,drug,drug>'
INN[INN=inn] -> '<INN,inn,inn>'
NUMERICVALUE[t=decint,ddosval=d_dos_val] -> '<DEC,d-dos-val,drug-dosage-value>'
NUMERICVALUE[t=decint,ddosval=d_dos_val] -> '<INT,d-dos-val,drug-dosage-value>'
ABSORPTION[A=A] -> '<"a",O,O> <"liberation",O,O> <ABSORPTION,A,absorption>'
```

FIGURE 3.8 – Langage des prescriptions représenté par la grammaire à traits utilisée pour la génération de données

L'extrait de règles présentées sur la Figure 3.8 définissent seulement un médicament et son identification. La première règle définit la règle de haut niveau permettant de générer le langage engendré par la suite. Les deux règles suivantes définissent l'acte de langage sur le médicament comme étant soit une spécialité ou une spécialité suivie de sa DCI. Ensuite les règles (4) et (5) définissent l'acte de dialogue permettant d'identifier le médicament. La première représentant le dosage du médicament et son unité ou sa concentration en pourcentage et la deuxième pour exprimer des médicaments comme « Tromadol dosé à 100 mg à libération prolongée ». La règle (6) permet de définir la règle pour construire le dosage, l'unité du médicament et (7) le dosage en pourcentage.

Dans notre méthode, on représente les terminaux sous forme de triplets adaptés pour la tâche de compréhension automatique du langage. Un triplet est composé de *<valeur,étiquette,étiquette-valeur>*. Cette représentation est la même que celle définie au début de la section 3.4. Sur la Figure 3.8, les valeurs sont représentées soit par des mots clés écrits tout en majuscules (DEC, INT, etc...) soit entre les guillemets qui représentent des chaînes de caractères. La génération d'une « phrase(S) » dénote une prescription composée d'un enchaînement de triplets avec des mots clés. Ensuite, ces mots clés en concordance avec les étiquettes valeurs sont remplacés aléatoirement par des vrais de vraies valeurs. Ce remplacement est fait en dehors de la génération des triplets et à posteriori de la génération du texte. L'exemple (1) montre un exemple d'ensemble de triplets composant un médicament et son identification ainsi que le remplissage avec des valeurs remplies aléatoirement

sur l'exemple (2).

(1) <INN,inn><"générique",g><"à",O><"libération",A><ABSORPTION,A><"dosé",O><"à",O>
<DECINT,d-dos-val><"%",dd-dos-up><"solution",d-dos-form><"injectable",d-dos-form-ext>

(2) Paracétamol générique à libération prolongée dosé à 0,0025 solution injectable

Comme sur l'exemple (2), une grammaire de génération pourrait générer des chaînes définies dans la grammaire dont le nombre d'exemples concaténés est quasiment infini. En revanche, le remplacement de valeurs et l'enchaînement de ces informations se feront sans aucune connaissance métier en s'éloignant des exemples du domaine. Cette grammaire pourrait produire une prescription absurde comme dans l'exemple (2), du paracétamol administré par la voie intraveineuse, alors que celui-ci se présente sous forme de gélules vu qu'il n'y a aucune contrainte sur la génération. Pour pallier à ce problème, nous avons contraint le remplissage à des valeurs déterminées à partir du thésaurus médical Thésorimed©. Le DCI et les spécialités sont choisis de façon aléatoire mais une fois que le médicament est identifié, les autres attributs sont choisis en limitant le nombre de valeurs possibles au niveau de la voie d'administration, la forme, etc.

La génération permet d'obtenir des données, mais ne permet pas d'obtenir une répartition équilibrée de classes. A cet effet, nous avons ajusté la technique de génération pour prendre en compte la répartition des attributs et la taille variable des prescriptions. De façon contrôlée, dans chaque itération, les classes les moins représentées sont choisies comme candidates potentielles à générer jusqu'à ce que la quantité d'attributs en général soient proche. Pour avoir des prescriptions comportant plus ou moins d'informations, le nombre de classes représentées par la génération est aussi choisi aléatoirement(5 à 10 attributs aléatoires). Le tableau suivant résume cette méthode de génération avec des exemples de candidats à générer et les exemples produits par la grammaire 3.4.

3.5 Intégration du système de dialogue avec un LAP

Nous avons précédemment décrit notre approche du dialogue dans le domaine des prescriptions médicamenteuses dans la section 3.3.2. Une partie très importante de cette modélisation dépend de l'interaction avec la connaissance experte d'un LAP (connaissance métier) et d'un thésaurus médical. La haute autorité de santé (HAS) définit un LAP comme étant « un logiciel dont au moins une des fonctions permet d'élaborer et d'éditer les prescriptions médicales. Il existe deux types de LAP : le LAP hospitalier, utilisé en établissement de santé et qui permet de limiter les choix de médicaments à des listes définies (livret thérapeutique) et le LAP de médecine ambulatoire »[2]. Dans notre cas, nous nous sommes basés sur un LAP certifié par la HAS dans un contexte hospitalier : Futura Smart Design©. Les LAP comportent beaucoup de fonctionnalités qui dépendent du contexte d'utilisation hospitalier ou ambulatoire. En revanche, les contrôles nécessaires pour les prescriptions médicamenteuses s'appliquent dans les deux contextes.

2. https://www.has-sante.fr/jcms/c_989142/fr/certification-des-logiciels-des-professionnels-de-sante

Candidats	Phrase (Prescription)
['inn', 'A', 'dos-val', 'dos-uf', 'min-gap-val', 'min-gap-ut', 'rhythm-perday', 'freq-startday']	valproate de sodium lp 7 comprimés par jour mini 25 minutes entre chaque prise 3 fois par jour à partir de ce midi
['inn', 'dos-val', 'dos-uf', 'rhythm-hour', 'dur-val', 'dur-ut']	lopéramide 4 gélules par jour à 6 heures pour 10 mois
['inn', 'dos-val', 'dos-uf', 'dos-cond', 'cma-event', 'fasting', 'freq-startday', 're-val', 're-ut', 'nr']	simvastatine 11 comprimés par jour en cas de constipation au cours de chacun des 3 repas à jeun à partir de demain traitement à renouveler pendant 83 jours non remboursable
['drug', 'd-dos-form', 'd-dos-form-ext', 'dos-val', 'dos-uf', 'nr']	metformine arrow comprimé pelliculé 4 comprimés par jour non remboursable
['inn', 'G', 'dos-val', 'dos-uf', 'min-gap-val', 'min-gap-ut']	lisinopril dihydraté générique 9 comprimé par jour mini 93 minutes entre chaque prise
['drug', 'roa', 'dos-val', 'dos-uf', 'min-gap-val', 'min-gap-ut', 're-val', 're-ut']	levocetirizine en orale 5 comprimés par jour minimum 30 secondes entre chaque prise traitement à renouveler pendant 18 jours
[drug', 'd-dos-val', 'd-dos-up', 'd-dos-form', 'd-dos-form-ext', 'dos-val', 'dos-uf', 'nr', 'qsp-val', 'qsp-ut']	grisefuline 250 mg comprimés sécables 1 comprimé non remboursable quantité suffisante pour 22 mois
['inn', 'G', 'd-dos-form', 'd-dos-form-ext', 'dos-val', 'dos-uf', 'fasting', 'rhythm-perday', 'ns']	gadotéridol géné solution injectable 10s solution à jeûn 3 fois par jour non substituable

TABLE 3.4 – Exemples de prescriptions produites par la grammaire à traits à partir des candidats d'attributs

Les contrôles sur les prescriptions sont effectués afin d'améliorer la qualité et la sécurité des prescriptions et rentrent dans la lutte contre la iatrogénie médicamenteuse notamment chez les personnes âgées de plus de 65 ans. Enjeu de santé publique, l'iatrogénie médicamenteuse est définie selon L'Assurance Maladie (Ameli) de la manière suivante : « La iatrogénie médicamenteuse désigne l'ensemble des effets indésirables provoqués par la prise d'un ou plusieurs médicaments. » La iatrogénie médicamenteuse est toutefois souvent évitable et les LAP font partie des outils utilisés dans cette démarche notamment chez le sujet âgé. En effet plus le nombre de médicaments à prendre est important (polymédication) plus le risque iatrogène augmente. Les contrôles de sécurité nécessaires des prescriptions médicamenteuses sont les suivantes selon la HAS :

— Redondance des substances actives : Le système alerte le prescripteur lorsqu'il y a un autre médicament ayant la même substance active parmi les traitements en cours du patient.

— Allergie, hypersensibilité ou l'intolérence au médicament prescrit

— Contre-indication par l'âge, le sexe, les antécédents ou les états physiopathologiques

— L'interaction du médicament avec une grossesse ou dans le cas d'un homme en âge de procréer, risques en cas de procréation

— Quantité maximale prescrite pour 24 heures : calcul du cumul du dosage afin d'assurer qu'il ne dépasse pas le seuil maximal pour 24 heures

— Vérification de l'intervalle minimale entre les deux prises d'un même médicament

— Vérification du dosage maximale par prise

— Vérification de la durée minimale et maximale de la prescription

— Incompatibilités physico-chimiques spécifiques au mélange de médicaments parentéraux ou topiques

— Vérification de l'impact sur la conduite de véhicules

— Vérification de si le médicament comporte un caractère dopant

— Dans le cadre d'une affection de longue durée (ALD) exonérante, indication sur les médicaments qui ne sont pas pris en charge

— Signalement des médicaments dont la prescription est restreinte à certains médecins spécialistes

Les contrôles de sécurité cités ci-dessus doivent être obligatoirement effectués par les LAP certifiés par la HAS. Il est à noter que ces contrôles sont à titre indicatif. C'est-à-dire que les médecins en prenant connaissance des ces *alertes*, peuvent prescrire une prescription même si elle est contre-indiquée. Cependant, suivant les réglementations de HAS, les prescripteurs doivent obligatoirement indiquer une motivation à leurs prescriptions. Le fonctionnement correct de ces systèmes complexes dépend de plusieurs acteurs : les pharmaciens qui paramètrent les unités de prescriptions et les informations concernant les médicaments, les personnels qui constituent et qui mettent à jour les dossiers électroniques des patients et les spécialistes qui effectuent leurs diagnostiques et les prescriptions sur les LAP.

Dans l'intégration de cette connaissance métier dans un contexte dialogique, nous avons fait le choix de découpler l'identification du médicament du reste des vérifications à faire en lien avec le dossier des patients. De même, nous avons séparé la modélisation concernant l'identification du médicament sans prendre en compte le livret thérapeutique pour mettre en place une méthodologie applicable dans un usage hospitalier et ambulatoire. La méthodologie de l'intégration avec un LAP pourrait être vue comme une extension à l'approche que nous avons décrite dans 3.3.2. Notre proposition sous forme d'extension à l'approche au dialogue que nous avons présenté précédemment est schématisée sur la Figure 3.9.

Notre approche de la modélisation d'une intégration avec un LAP est étroitement liée au référentiel fonctionnel de certification des logiciels d'aide à la prescription qui est publié et revu par la HAS. L'étape **(1)** de l'intégration porte sur l'identification du médicament. Dans le cas d'une prescription dans un milieu hospitalier, le choix des médicaments prend en compte les médicaments paramétrés dans le livret thérapeutique. Nous identifions également s'il s'agit d'une prescription en DCI dans cette étape. L'étape **(2)** permet d'identifier la voie d'administration de la posologie. Dans cette étape, on distingue la voie d'administration du médicament de la voie d'administration de la posologie. En effet, un médecin peut demander la prise d'un médicament par une voie d'administration différente que celle pour laquelle il a été commercialisé. Un exemple pourrait être des gouttes ophtalmiques administrées oralement. La prise en compte des aspects liés à la voie d'administration est au paramétrage des médicaments par les pharmaciens d'officines.

La même chose s'applique sur les unités de prescription. L'étape **(3)** montre cette logique permettant d'identifier une forme galénique paramétrée pour le médicament en question. Pour des désignations vagues comme « 1 injection le matin », cette étape reformule les unités

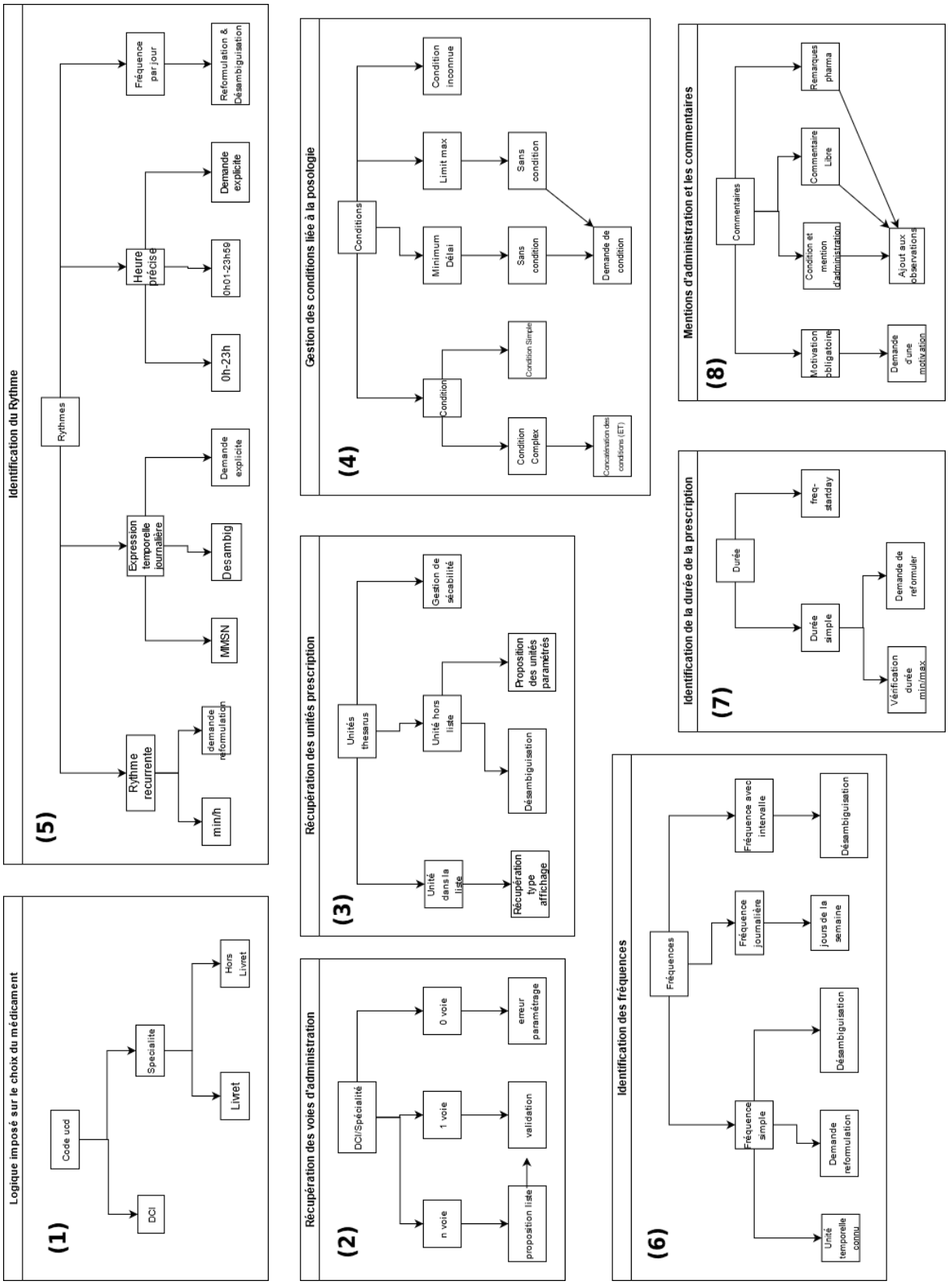


FIGURE 3.9 – Extension de notre approche au dialogue pour une intégration avec un LAP

dites vagues sous des formes galéniques par défaut qui permettent aux prescripteurs d'être plus précis dans leurs formulations. Nous effectuons également les vérifications liées à la gestion de sécabilité des médicaments à conditions que les paramètres de sécabilité du médicaments aient bien été renseignés dans le logiciel par les pharmaciens. Par exemple, le système devrait avertir le prescripteur lors de la prescription d'un demi comprimé d'un médicament non sécable.

L'étape **(4)** englobe la gestion de conditions et les délais à respecter pour la prise d'un médicament. A l'oral et sur les exemples que nous avons extraits à partir de Lariven (2008), la formulation d'une prescription comme « 1 à 2 comprimés » est insuffisamment précise pour la validité d'une prescription pour un LAP. Les prescriptions en anglais sont formulées d'une façon beaucoup plus codée identifiant les rythmes en utilisant les fréquences comme *q.d*(1 fois par jour), *b.i.d.*(2 fois par jour), *t.i.d.*(3 fois par jour) et *q_xh*(x fois par 24 heures). En français, la formulation du rythme et de la fréquence d'une prescription est beaucoup moins codée qu'en anglais. Dans Futura Smart Design©, il est obligatoire de préciser un délai minimal entre les deux prises ou suivant *q_xh*, toutes les tant heures ou minutes afin d'éviter les ambiguïtés.

Les étapes **(5)** et **(6)** sont consacrées à l'identification du rythme et la fréquence de la prescription. Cette identification est très proche de notre modélisation sémantique du domaine. La seule extension porte sur les désambiguïsations nécessaires. De la même façon, l'étape **(7)** suit la même logique concernant l'identification de la durée. En revanche, dans une langue naturelle, notre sémantique couvre des informations qui ne peuvent pas rentrer dans un schéma posologique classique. Par exemple, les conditions et mentions d'administration comme « après le petit-déjeuner » ou « à jeun » ou les remarques pharmaceutiques sont convertis en commentaires. Cette dernière étape est schématisée dans l'étape **(8)** de la figure 3.9.

3.6 Méthodes d'évaluation

Dans le cadre d'un système de dialogue orienté but, il est crucial d'estimer la performance du système. Même si on peut extraire des statistiques du système global ou de ses sous composants, un dialogue réussi dépend largement de l'application en question (Deriu et coll., 2021). Dans le contexte d'un système de dialogue modulaire, il faut mesurer la performance de la globalité du système de dialogue ainsi que la performance des composants tels que la compréhension du langage (NLU en anglais), la reconnaissance automatique de la parole (RAP), la synthèse vocale (TTS en anglais), etc. Dans cette section nous allons présenter les mesures que nous avons utilisées dans l'évaluation du système de dialogue et ses composants.

3.6.1 Évaluation Globale du système de dialogue

La mesure la plus connue pour estimer la qualité d'un système de dialogue orienté but dans sa globalité est de calculer le taux de réussite de l'achèvement de la tâche (task success rate). Cette mesure permet d'évaluer tout simplement à la fin du dialogue si la tâche est correctement effectuée ou non. Dans notre système, la tâche étant d'effectuer une prescription médicamenteuse, le taux de réussite de la tâche indique si une prescription a été ajoutée sur le dossier du patient suite à la validation du prescripteur. En revanche, vu que nous sommes sur une tâche complexe destinée aux experts du milieu médical (prescripteurs), différents niveaux de granularité de la définition de la même tâche sont envisageables. Par exemple, nous pouvons considérer que la tâche est réussie si la prescription est acceptée par le prescripteur et a été validée par le LAP. Une granularité plus fine pourrait être, étant donné le profil d'un patient fictif, de considérer comme réussie une prescription à l'oral d'un médicament en vérifiant la validité par le prescripteur, le LAP et un expert médical vérifiant manuellement les éventuelles contre-indications. Dans cette thèse, nous nous sommes limités à la validité sémantique : la validation par le prescripteur de la prescription affichée suite à l'interaction avec le système.

Dans un système de dialogue orienté but, le taux de réussite de la tâche donne une indication sur la réussite globale de la tâche. Pour avoir une idée plus fine de la satisfaction des utilisateurs, l'une des pratiques courantes est de faire compléter un questionnaire de satisfaction aux utilisateurs suite à leur échange avec le système (Jurafsky et Martin, 2014). Cette mesure repose sur l'idée qu'il serait possible d'estimer l'utilisabilité d'un système par la satisfaction de ses utilisateurs (Deriu et coll., 2021). Cette mesure permet de calculer un taux de satisfaction globale de l'utilisateur et d'évaluer l'efficacité et la pertinence du système en répondant aux questions spécifiques en utilisant une échelle de Likert : une échelle allant de 1 à 5. Dans cette échelle 1 indique l'état d'accord le plus fort et 5 indique l'état de désaccord le plus fort. La valeur moyenne des réponses à toutes les questions donne la valeur de satisfaction globale d'après cette méthode.

Dans la conception des questionnaires de satisfaction, nous nous sommes inspirés de Walker et coll. (2000) et Dzikovska et coll. (2011). Nous avons modélisé 18 questions pour les utilisateurs naïfs et 6 questions supplémentaires pour les experts médicaux. Ces questions sont catégorisées dans les rubriques suivantes : évaluation de l'échange interactionnel, l'évaluation globale, le protocole d'évaluation et les questions sur la pertinence du système (pour les experts médicaux). Les questions sur l'évaluation du système portent plus au niveau des dialogues produits et sur la qualité et la pertinence du système à donner des réponses adéquates. Les questions sur la globalité du système portent sur la satisfaction globale de l'échange avec le système. La troisième partie des questions n'est pas techniquement orientée vers le système de dialogue mais sur la compréhension et l'adéquation de la tâche demandée dans l'expérience. En dernier, pour les experts médicaux, nous avons inclus quelques questions afin d'évaluer la pertinence d'effectuer des prescriptions médicamenteuses à l'oral par rapport à la pratique courante en médecine.

Le calcul du taux de satisfaction à partir des réponses à un questionnaire pourrait être accompagné de mesures supplémentaires. Ces mesures supplémentaires peuvent être plus ou moins détaillées, au niveau de l'échange ou sur la globalité. Par exemple, une mesure au niveau global que nous pourrons pouvons extraire automatiquement est le nombre de tours de parole écoulés pour arriver à l'achèvement de la tâche. Cette mesure est d'autant plus importante pour nous afin de comparer le temps de saisie des prescriptions orales par rapport à la rédaction d'une ordonnance ou une saisie manuelle de prescription médicamenteuse. Au niveau de l'échange, des mesures comme le taux de mots reconnus correctement par le système de reconnaissance de la parole ou le taux d'attributs extraits correctement par le remplissage d'attributs peut permettre de calculer une corrélation entre ces mesures et la satisfaction de l'utilisateur. Ces métriques seront détaillées dans la partie 3.6.2.

Un autre aspect important concernant l'évaluation des systèmes de dialogue porte sur le déroulement et le déclenchement du dialogue. Pour des raisons de confidentialité, dans notre système de dialogue, l'échange est déclenché à l'initiative de l'utilisateur avec un principe de *push-to-talk* (appuyer-pour-parler). De même, pour la collecte de données la parole est coupée en cliquant sur le bouton « arrêter » comme dans un enregistrement. Certaines mesures permettant de mesurer la frustration de l'utilisateur comme les *barge-in* (la coupure du système par l'utilisateur), le nombre d'annulation de la tâche ou le nombre de demande d'aide pour effectuer une tâche nous donnent la possibilité d'estimer la qualité de l'échange. Dans notre approche, nous avons gardé seulement le nombre d'annulation d'une tâche.

3.6.2 Évaluation des modules du système de dialogue

Dans l'architecture modulaire, le premier module est celui de la reconnaissance automatique de la parole (RAP). Pour la RAP, nous avons utilisé la RAP des services de Google. Dans cette thèse, la performance du service de RAP est calculée par le taux d'erreur de mots ou *Word Error Rate* (WER) en anglais. Le WER est une mesure classique dérivée de la distance de Levenshtein et consiste à calculer le nombre d'insertions (I), de substitutions (S) et délétions (D) par rapport à une transcription experte de référence composée d'une séquence de N mots. Pour calculer cette mesure, la séquence de mots reconnus (hypothèse) est d'abord alignée avec la séquence de mots de la transcription experte (référence) par une technique de programmation dynamique. Puis le WER est calculé comme suivant :

$$WER = \frac{I + S + D}{N} \tag{3.1}$$

Dans cette mesure, plus le WER est bas, plus l'hypothèse est proche de la transcription de référence. De façon similaire, une mesure inspirée de WER mais qui évalue la performance de la compréhension automatique du langage est la mesure de taux d'erreur de concepts ou concept error rate en anglais (CER). Le CER classique est calculé de la même manière sauf que les séquences ne sont pas les séquences de mots mais sont des séquences d'attributs ou de concepts. Dans ce travail, nous utilisons une extension de cette mesure pour les séquences non alignées proposée par Mishakova et coll. (2019a). Dans cette extension, la sé-

quence est triée alphabétiquement avant d'effectuer l'alignement dynamique et le calcul de taux d'insertion/délétion/substitution. De cette manière, un taux de CER plus proche de la réalité d'une chaîne générée par des méthodes génératives comme seq2seq.

Dans cette thèse, nous utilisons aussi les mesures plus classiques utilisées dans le domaine de la compréhension. Dans la tâche de compréhension, à partir d'un énoncé, on extrait l'intention du locuteur est extraite et les valeurs sont remplies dans les cadres sémantiques. La tâche de détection d'intention est une tâche de classification binaire. Pour évaluer sa performance nous avons utilisé les mesures classiques de précision, rappel et f-mesure : précision= $\frac{VP}{VP+FP}$ et rappel=$\frac{VP}{VP+FN}$. Dans cette mesure, nous évaluons les prédictions du système par rapport aux classes de référence en distinguant les vrais positifs (VP), les vrais négatifs (VN), les faux positifs (FP) et les faux négatifs (FN). La *précision* est la proportion des classes prédites qui sont correctes sur l'ensemble des prédictions. Le *rappel* est le taux des classes prédites qui sont correctes sur l'ensemble des classes à prédire. Vu que la précision donne une estimation sur l'exactitude des prédictions tandis que le rappel estime une mesure d'exhaustivité, nous calculons également la f-mesure qui est une moyenne harmonique de précision et de rappel comme comme ce qui suit :

$$F1 = \frac{2 * précision * rappel}{précision + rappel} \tag{3.2}$$

Nous abordons la tâche de compréhension des concepts comme une tâche d'étiquetage de séquences. La performance peut-être calculée de façon similaire à la classification binaire en utilisant la précision, le rappel et la f-mesure. Pour cette tâche de classification, chaque mot (*token*) doit recevoir une étiquette de classe. En revanche, différents niveaux de granularité peuvent être pris en compte dans l'attribution de ces classes. Pour les concepts, nous distinguons l'étiquette, l'étiquette-valeur et la valeur. Les étiquette-valeurs peuvent être appelées aussi des valeurs normalisées. Nous pouvons calculer ces mesures sur l'attribution correcte des étiquettes ou sur les valeurs normalisées. Pour calculer la performance du système, il est important de calculer la précision, le rappel et la f-mesure individuellement pour chaque cadre mais aussi de façon globale en effectuant une agrégation de résultats. La première méthode d'agrégation de résultats consiste à calculer la **micro-moyenne**. Similaire à la classification binaire, le calcul de précision est calculé comme la somme des vrais positifs (VP) de toutes les classes parmi toutes les prédictions positives.

$$Précision(Micro) = \frac{VP_1 + VP_2 + ... + VP_n}{VP_1 + VP_2 + ... + VP_n + VP_1 + FP_2 + ... + FP_n} \tag{3.3}$$

De façon similaire, la micro-moyenne de rappel est calculée de la façon suivante :

$$Rappel(Micro) = \frac{VP_1 + VP_2 + ... + VP_n}{VP_1 + VP_2 + ... + VP_n + FN_1 + FN_2 + ... + FN_n} \tag{3.4}$$

La deuxième méthode d'agrégation consiste à calculer la **macro-moyenne**. Pour calculer la macro-moyenne des prédictions, on calcule la moyenne de la précision et du rappel ce qui suit : précision=$\frac{préc_1 + préc_2 + ... + préc_n}{n}$, rappel = $\frac{rappel_1 + rappel_2 + ... + rappel_n}{n}$. L'évaluation au

niveau micro permet de calculer une somme sur les valeurs les plus fines puis de calculer la mesure en prenant en compte chaque donnée tandis que dans macro-moyenne on calcule les mesures puis on prend les moyennes.

Pour résumer, le tableau 3.5 fait une synthèse des mesures utilisées dans cette thèse pour calculer la performance du système de dialogue, le module de RAP et de NLU.

Mesure	**Module**	**Mesure**	**Module**
WER	RAP	Le taux de réussite de la tâche	Dialogue
CER	NLU	Moyenne de tours de dialogue	Dialogue
Précision	NLU	Temps écoulé	Dialogue
Rappel	NLU	Taux d'erreurs système (*bug*)	Dialogue
F-mesure	NLU	Taux d'annulation ou reccomencer	Dialogue

TABLE 3.5 – Synthèse des mesures utilisées dans l'évaluation des performances

CHAPITRE 4

Compréhension automatique des prescriptions médicamenteuses

Dans ce chapitre, nous allons détailler l'approche de compréhension automatique sur les transcriptions textuelles de prescriptions médicamenteuses dans un contexte de ressources limitées.

Bien que le domaine des prescriptions médicamenteuses soit très normé, l'analyse automatique de celles-ci est rendue difficile par la variabilité du contenu des prescriptions médicamenteuses. Ainsi, la modélisation du domaine présentée dans 3.3 dépend largement des caractéristiques des prescriptions. Dans 3.3, nous les avons présentés de manière générale. En ce qui concerne la compréhension des prescriptions, nous avons fait une analyse plus détaillée regroupant ces différents cas de prescriptions. Cette analyse est détaillée dans l'annexe A. Cette section aborde également les contraintes posées pour les LAP par la HAS pour sécuriser les prescriptions électroniques que nous avons abordées succinctement en sections 3.3 et 3.5.

La section 4.1 présente la manière dont nous avons constitué un premier corpus pour l'analyse des prescriptions et le développement des modèles de compréhension. Nous détaillons les techniques mises en place pour faire face au faible nombre de données d'apprentissage et au déséquilibre de représentation des différentes classes.

Ensuite, la section 4.2 présente les architectures des modèles de compréhension que nous avons acquis par apprentissage automatique. Nous abordons la compréhension sous forme de remplissage d'attributs (*slot-filling*). Dans cette section nous étudions plusieurs modèles de type séquence à séquence.

Étant donné le contexte de faible de ressources dans lequel s'est déroulée la thèse, nous nous sommes intéressés à des approches utilisant : des techniques d'augmentation de données, des techniques semi-supervisées (pour tirer parti de données non annotées) et des techniques utilisant des modèles pré-entraînés. la section 4.3 détaille donc toutes les expériences associées à cette étude.

4.1 Constitution du corpus de prescriptions

Face au paradoxe de l'œuf et de la poule bien connu dans la modélisation dialogique des interactions nouvelles avec un agent (Fraser et Gilbert, 1991), nous avons suivi une approche

itérative pour construire un corpus du domaine. Cette approche itérative ainsi que son calendrier est détaillée dans la section 3.3.4 du manuscrit. Le corpus de NLU que nous avons constitué a donc évolué au fil de cette thèse. La section 4.1.1 présentera l'étape de la constitution initiale de ce corpus qui nous a permis d'étudier les prescriptions médicamenteuses pour définir un domaine, d'initier (*bootstrap*) des modèles de compréhension et constituer un premier ensemble de données utilisées pour apprendre et évaluer des modèles.

4.1.1 Définition des attributs et premier corpus

L'approche que nous avons adoptée est une approche de remplissage d'attributs (*slot-filling*) qui se base sur la définition des cadres qui découle de la théorie introduite par Fillmore et coll. (1976). La méthode de définition des cadres et leurs catégories de façon itératives en consultant un expert du domaine biomédical à partir de Lariven (2008) est détaillé dans la section 3.3.1. Dans cette partie, nous allons présenter les attributs qui définissent le domaine des prescriptions médicamenteuses. Le Tableau 4.1 présente ces attributs et leurs définitions avec un exemple.

Catégories	Cadres	Définitions	Exemples
Médicament	drug	Le nom de spécialité ou produit médicamenteux en vente libre	Doliprane©
	inn	La molécule, la dénomination commune internationale (DCI)	Paracétamol
	G	Si le médicament est générique	générique
Identification du Médicament	d-dos-val	La valeur numérique du dosage de l'intensité du médicament	500
	d-dos-up	L'unité de mesure du dosage ou degré de pourcentage de concentration	milligrammes
	d-dos-form	La forme galénique du médicament	comprimés
	d-dos-form-ext	Spécifications liées à la forme galénique (ex. effervecent, bisécable)	effervescents
	A	Caractéristiques sur l'absorption d'un médicament	à libération **prolongée**
	roa	La route d'administration du médicament	voie **orale**
Dosage et les Instructions	dos-val	L'unité du dosage de la posologie destiné au patient	3
	dos-uf	L'unité de mesure (mg,kg,...) ou une forme (cpr, patch) spécifié	cachets
	dos-cond	Condition de la prise (maux de tête, insomnie,...)	en cas de **douleurs**
	max-unit-val	Le nombre d'unités à ne pas dépasser	à ne pas dépasser **4**
	max-unit-uf	L'unité ou forme liée à l'unité de mesure	comprimés
	max-unit-ut	L'intervalle de temps à respecter	par **jour**
	min-gap-val	La valeur numérique de l'espacement minimal entre les prises	espacés de **15**
	min-gap-ut	L'unité de temps précisé pour l'espacement minimal	**mins** entre chaque prise
	cma-event	Conditions ou mentions d'administration (ex. avec un verre d'eau)	avec un verre d'eau
	fasting	S'il faut prendre le médicament à jeun ou non	à jeun
Fréquence et Rythme	rhythm-rec-val	La valeur numérique d'un rythme récurrente (ex. chaque 2 heures)	toutes les 4
	rhythm-rec-ut	L'unité de temps associé au rythme récurrente	heures
	rhythm-tdte	Le temps de la prise spécifié au sein d'un jour	**matin** et **midi**
	rhythm-hour	L'heure de la prise	à **14** heures
	rhythm-perday	Nombre de prise par jour	**3** fois par jour
	freq-val	Valeur numérique de la fréquence de la prise	tous les **2**
	freq-ut	Unité du temps associé à la fréquence de la prise	jours
	freq-startday	Spécification du début de la prescription	à commencer **demain**
	freq-days	Les jours de la prise	les **lundis** et les **mardis**
	freq-int-v1	La valeur numérique du début d'un intervalle	à prendre **2**
	freq-int-v1-ut	L'unité du temps du début d'un intervalle	jours
	freq-int-v2	La valeur numérique de la fin d'un intervalle	sur **4**
	freq-int-v2-ut	L'unité du temps de la fin d'un intervalle	semaines
Durée	dur-val	L'unité numérique de la durée exprimée	pendant **6**
	dur-ut	L'unité de temps de la durée	mois
Remarques Pharma	re-val	La valeur numérique si la prescription est à renouveler	à renouveler pendant **4**
	re-ut	L'unité du temps du renouvellement	mois
	ns	Si le médicament ne doit pas être subsituté	non substituable
	nr	Si le médicament est non remboursable	non remboursable
	qsp-val	La valeur numérique du temps exprimé en quantité suffisante pour	quantité suff. pour **3**
	qsp-ut	L'unité du temps d'un temps exprimé en quantité suffisante pour	semaines

TABLE 4.1 – Explication des cadres définis pour l'approche de remplissage d'attributs qui définit l'espace sémantique des prescriptions médicamenteuses

Le tableau récapitulatif 4.1 présente l'ensemble des attributs que nous avons retenu à l'issue de l'approche itérative. Nous avons utilisé des méta-catégories (qui ne font pas partie du système) pour regrouper les attributs liés entre eux. Certains attributs sont fortement corrélés. Par exemple, la route d'administration par défaut d'un médicament est liée à sa forme galénique. Cependant, dans certains cas, un prescripteur peut demander la prise d'un médicament par une voie d'administration autre que celle par défaut (p. ex. un comprimé écrasé et administré par voie intraveineuse). Ceci explique qu'il existe deux attributs séparés. Les autres catégories concernent globalement la posologie. Selon le dictionnaire de L'Académie Nationale de Pharmacie la posologie est définie comme suivante :

> « Étude et recommandation des doses efficaces et bien tolérées des substances médicamenteuses en fonction de leur nature, de leur activité et des caractéristiques physiologiques (âge, sexe, poids, habitudes hygiénodiététiques) et pathologiques éventuelles du sujet traité. »

Suivant cette définition classique d'une posologie, nous avons constitué les méta-catégories figurant sur le Tableau 4.1. Cependant, nous associons souvent la durée d'une prescription à la posologie. En réalité, l'administration d'un médicament fait partie plutôt d'un schéma posologique. L'exemple suivant illustre une prescription qui suit un schéma posologique dégressif.

- Cotrancyl© 20 mg, 3 comprimés le matin pendant 3 jours puis 2 comprimés par jour pendant 5 jours puis 1 comprimé par jour pendant 5 jours puis arrêt.

La prescription ci-dessus montre bien cette notion dégressive des doses qui constitue la posologie. Nous remarquons que la posologie est composée de différentes doses, temps de prises et durées qui diminuent au fur et à mesure avant d'être complètement arrêtés. Cela nous fait revenir sur la notion de posologie composée de quatre volets : la voie d'administration, la dose, le rythme, la fréquence de prise et la durée du traitement. Dans cette définition plus large, nous avons fait le choix de décomposer les temps de prises en « rythme » et « fréquence » pour prendre en compte les schémas posologiques différents.

Il y a donc une partie concernant le rythme de la prescription où l'on précise le nombre de prises quotidiennes et une fréquence qui permet de définir l'intervalle de répétition de ces prises. Par exemple, « 1 comprimé toutes les 4 heures (le rythme) tous les 2 jours (la fréquence) ». Il est à noter que les conditions et mentions d'administration peuvent comporter des précisions liées au rythme et la fréquence d'une posologie (avant la douche, pendant le petit-déjeuner, etc.) que nous ne retenons pas dans la catégorie de rythme ou de fréquence. Toutefois, une prescription naturelle peut comporter un certain nombre d'informations qui doivent être déduites par un système automatique. Par exemple, lorsqu'un prescripteur exprime « ... 1 comprimé 1 jour sur 2 », le système doit déduire l'intervalle de temps de fin. La figure 4.1 schématise ce processus de remplissage d'attribut. 1 jour sur 2 fait bien évidemment référence à 1 jour sur 2 (jours) et non des semaines ou des mois lorsqu'il n'est pas exprimé.

MEDICAL_PRESCRIPTION(..., freq-int-v1 = 1,
freq-int-v1-ut = JOURS,
freq-int-v2 = 2,
freq-int-v2-ut = ?)

FIGURE 4.1 – Remplissage d'attributs avec une fréquence à intervalles

On retrouve cette notion même dans les prescriptions les plus simples. Par exemple, il est évident que lorsqu'un médicament est prescrit en précisant « tous les jours », on parle de tous les **1** (*freq-val*) **jours** (*freq-ut*). Comme expliqué dans 3.3.1, le fait de décomposer les cadres en unités et valeurs permet au LAP de contrôler de façon canonique la cohérence de tous les éléments. Par contre, cela nécessite une désambiguïsation préalable de toutes les valeurs.

drug	816	**max-unit-ut**	8	**freq-int-v1**	8	**freq-int-v1-ut**	8
inn	57	**min-gap-val**	0	**min-gap-ut**	0	**G**	6
freq-int-v2	8	**freq-int-v2-ut**	9	**d-dos-val**	518	**d-dos-up**	418
cma-event	69	**fasting**	2	**dur-val**	197	**dur-ut**	190
d-dos-form	126	**rhythm-rec-val**	5	**rhythm-rec-ut**	5	**d-dos-form-ext**	6
re-val	1	**re-ut**	1	**rhythm-tdte**	467	**A**	23
roa	71	**rhythm-hour**	10	**ns**	0	**nr**	0
dos-val	761	**dos-uf**	712	**rhythm-perday**	189	**dos-cond**	19
freq-val	4	**freq-ut**	20	**qsp-val**	40	**qsp-ut**	37
max-unit-val	17	**max-unit-uf**	11	**freq-startday**	0	**freq-days**	6

TABLE 4.2 – Répartition des étiquettes d'attributs de GuideCorpus

Afin d'obtenir un premier jeu de données, nous nous sommes tournés vers des livres pédagogiques destinés aux étudiants en médecine. Parmi les livres pédagogiques, « Le guide des premières ordonnances » (Lariven, 2008) était particulièrement intéressant car ce livre était concernait plusieurs spécialités telles que la psychiatrie, la nutrition, la pédiatrie, etc. De cette manière, il a été possible d'inclure des prescriptions de différentes spécialités en médecine, écrites par des spécialistes variées. Les étapes que nous avons suivies pour extraire les prescriptions à partir de Lariven (2008) a été présenté dans 3.4.1 du chapitre méthode.

Les prescriptions que nous avons extraites à partir de Lariven (2008) nous a permis de construire un corpus constitué de 832 exemples de prescriptions. Dans la suite de cette thèse nous nous référerons à ce corpus en utilisant le nom **G**uideCorpus. La répartition de la fréquence des attributs du GuideCorpus est présentée sur le tableau 4.2. Cette première extraction de données nous a permis d'obtenir des exemples pour initialiser notre système de NLU mais elle souffre d'une inégalité importante de répartition entre les attributs. Cette répartition sera présentée dans la section 4.1.3

Comme les échantillons ont été répartis aléatoirement, il y a certains attributs représentés uniquement sur le corpus d'entraînement ou ceux qui sont sous-représentés. Nous pouvons remarquer que certains attributs présentés dans le Tableau 4.2 n'ont aucun exemple dans le corpus. Cette situation est liée à l'inclusion des catégories sémantiques issues d'un LAP selon la méthode décrite dans 3.3.1. Face à cette représentation inégale des classes, nous avons choisi une méthode de génération artificielle de données.

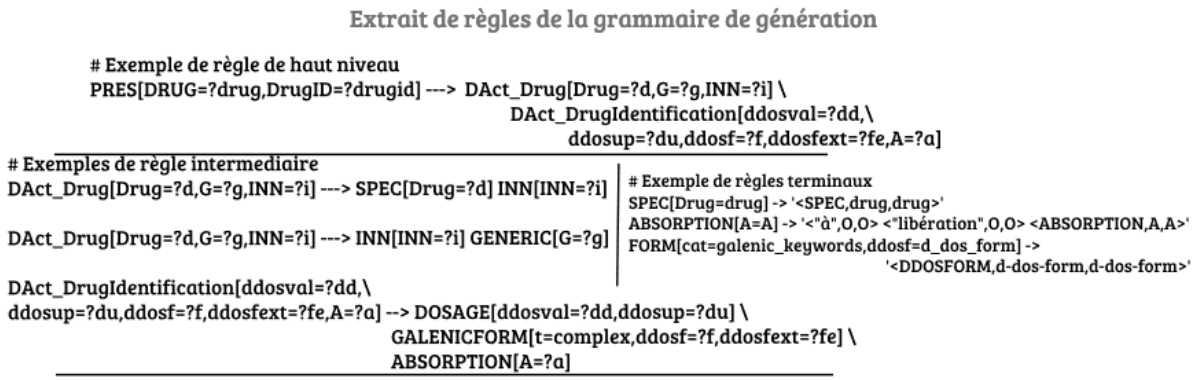
4.1.2 Génération des données artificielles

Nous avons choisi une technique permettant de générer des données de façon artificielle pour plusieurs raisons. Tout d'abord, les données figurant sur une prescription sont très variées et nécessiteraient beaucoup d'exemples pour un modèle de NLU stochastique de généraliser sur l'ensemble de ces informations. De même, pour un LAP, même un cas rare doit être représenté de la même manière qu'un cas standard par le système. Par ailleurs, dans une situation dialogique, seuls les exemples de prescriptions complètes ne suffiront pas, il faudrait des exemples de dialogue et de sous dialogues pour compléter/corriger/annuler les informations. Une fois définie, les règles de grammaire permettraient non seulement de générer des prescriptions complètes mais aussi celles incomplètes qui peuvent être utilisées pour compléter les informations dans l'approche de remplissage d'attributs.

Comme décrite dans 3.3.1, avec une grammaire, on peut représenter à la fois des attributs-étiquettes, attribut-valeurs et valeurs. Pour que les exemples soient cohérents, un système de génération devrait être en mesure d'injecter des connaissances externes dans la production des énoncés. Pour le domaine des prescriptions médicamenteuses, cette connaissance externe consisterait à remplir les noms de spécialistes, leurs DCI, leurs formes galéniques, etc. en fonction médicament en question. Nous avons présenté dans la partie 2.7 de l'état de l'art, quelques techniques de génération de données artificielles. L'une des approches classiques de la programmation logique (Prolog) permet d'incorporer ces connaissances facilement, mais rend la génération de phrases très difficile. Les grammaires hors-contexte, quant à elles, sont efficaces pour générer des prescriptions, mais ne peuvent pas intégrer facilement les caractéristiques requises. Les grammaires à base de traits sont un compromis entre les deux, permettant une génération facile des phrases, mais avec la possibilité d'attacher des caractéristiques aux phrases de production finale.

Malgré le manque de structures variées, cela nous a permis d'obtenir une quantité de données riches en information sémantique qui permet d'augmenter les données de prescription. La Figure 4.2 résume cette approche avec quelques règles et exemples de génération.

L'exemple de règle de haut niveau en haut de la Figure 4.2 représente une règle à base de traits qui définit une prescription incomplète contenant un médicament et l'information sur son identification. À gauche, les règles intermédiaires définissent qu'un médicament peut-être représenté par une spécialité ou une DCI. Concernant, l'identification du médicament, cet extrait montre un exemple composé d'un dosage, d'une forme galénique et d'un mode



Candidats	Phrase (Prescription)
['inn', 'A', 'dos-val', 'dos-uf', 'min-gap-val', 'min-gap-ut', 'rhythm-perday', 'freq-startday']	valproate de sodium lp 7 comprimés par jour mini 25 minutes entre chaque prise 3 fois par jour à partir de ce midi
['inn', 'dos-val', 'dos-uf', 'rhythm-hour', 'dur-val', 'dur-ut']	lopéramide 4 gélules par jour à 6 heures pour 10 mois
['inn', 'dos-val', 'dos-uf', 'dos-cond', 'cma-event', 'fasting', 'freq-startday', 're-val', 're-ut', 'nr']	simvastatine 11 comprimés par jour en cas de constipation au cours de chacun des 3 repas à jeun à partir de demain traitement à renouveler pendant 83 jours non rembour-sable

FIGURE 4.2 – Extrait de règles et d'exemples générés avec la grammaire hors-contexte à base de traits

d'absorption (ex. libération prolongée). À droite, on peut trouver des exemples de règles terminales qui génèrent les triplets d'attributs, d'attributs-valeurs et de valeurs. Il est à noter que la plupart des attributs sont déléxicalisés. Contrairement à une approche classique de définition de grammaire où les terminaux représentent le vocabulaire d'une langue engendrée par la grammaire, dans cette approche on produit des mots clés qui doivent être remplis par une connaissance experte. Le processus de la création des règles dans la conception de la grammaire artificielle a été élaborée dans la section 3.4.2.

Cette technique de génération produit un enchaînement d'informations sans prendre en compte une quelconque logique de contrôle de haut niveau sur la prescription d'un médicament. Par exemple, le système peut très bien concaténer une spécialité sous forme orale et continuer la génération comme s'il s'agissait d'une injection. De même, les valeurs numériques qui doivent être remplies sont remplacées par des chiffres aléatoires. En revanche, même si les valeurs numériques sont remplies aléatoirement par des chiffres, nous avons distingué les différentes catégories de valeurs numériques. Par exemple, les chiffres concernant le dosage d'un médicament sont de nature différente de celle des posologies. Par exemple, un médicament peut-être dosé à 0,025 milligrammes alors qu'une posologie ne pourrait jamais consister en 0,025 d'un comprimé. De même, les chiffres concernant les heures ne seraient pas les mêmes que ceux qui définissent les durées ou les intervalles. Bien évidemment, il serait très compliqué de concevoir des règles qui seraient capables de générer des prescriptions conformes suivant toutes les recommandations spécifiques selon les médicaments. En revanche, en utilisant une base de données médicamenteuses (BDM), au lieu de tout générer aléatoirement, un minimum de cohérence pourrait aider le système à générer des exemples plus pertinents. Pour contrôler la génération, nous avons utilisé la base de BDM de Thésorimed©(d'autres outils existent tel que Romedi proposé par Cossin et coll.

(2018) que nous avons découvert plus tardivement) .

En bas de la Figure 4.2, on voit un ensemble de candidats et des exemples de réalisation. Afin d'obtenir un corpus équilibré en termes de représentation des attributs, notre méthode de génération sélectionne les attributs candidats qui sont sous-représentés dans le corpus initial afin d'équilibrer en moyenne la plupart des attributs. Cependant, une prescription doit contenir un minimum d'information pour être cohérente. C'est pourquoi, certains attributs comme le dosage, le nom d'un médicament ou la durée sont sur-représentés. Le tableau 4.3 représente la répartition des attributs avant et après la génération artificielle des attributs. La génération de données a été appliquée uniquement dans la partie du corpus dédié à l'entraînement. La répartition finale du corpus en train, dev et test sera détaillée dans la sous-section 4.1.3.

Attributs fréquents	**Attributs rares**	**Attributs après équilibrage**	
drug : 416	inn : 29	drug :1631	inn : 1841
rhythm-perday :105	rhythm-rec-val : 2	rhythm-perday : 441	rhythm-rec-val : 446
d-dos-form :68	d-dos-form-ext : 3	d-dos-form : 944	d-dos-form-ext : 441
dur-val :108	re-val : 0	dur-val : 508	re-val :516

TABLE 4.3 – Distribution des attributs avant et après équilibrage des classes (sur les données d'entraînement)

Le changement des attributs montré sur le Tableau 4.3 que la globalité des attributs se sont équilibrés entre 400-550 exemples. Vu que certains attributs tels que *drug,dos-val,d-dos-form*, etc. sont nécessaires pour que la prescription soit compréhensible, ces attributs sont quand même sur représentés par rapport à d'autres attributs. Le Tableau 4.4, montre le nombre d'attributs du corpus d'entraînement après le processus de génération.

drug	1631	**roa**	458	**min-gap-ut**	440	**freq-val**	441	**dur-val**	508
inn	1841	**dos-val**	3414	**cma-event**	441	**freq-ut**	887	**dur-ut**	503
G	442	**dos-uf**	3391	**fasting**	441	**freq-startday**	440	**re-val**	516
d-dos-val	726	**dos-cond**	503	**rhythm-rec-val**	446	**freq-days**	888	**re-ut**	516
d-dos-up	447	**max-unit-val**	898	**rhythm-rec-ut**	446	**freq-int-v1**	442	**ns**	440
d-dos-form	944	**max-unit-uf**	450	**rhythm-tdte**	462	**freq-int-v1-ut**	442	**nr**	442
d-dos-form-ext	441	**max-unit-ut**	449	**rhytm-hour**	441	**freq-int-v2**	440	**qsp-val**	444
A	455	**min-gap-val**	440	**rhythm-perday**	441	**freq-int-v2-ut**	441	**qsp-ut**	444

TABLE 4.4 – Distribution de nombre de classes du corpus d'entraînement après l'équilibrage

4.1.3 Répartition finale du corpus de compréhension

Nous avons présenté les caractéristiques du GuideCorpus que nous avons constitué ainsi que la distribution des attributs avant et après la méthode de génération que nous avons mise en place. Cependant, les systèmes de NLU effectuent principalement deux tâches : non seulement le **remplissage d'attributs** mais également l'extraction de l'**intention** du locuteur. Pour l'intention, nous nous sommes préalablement limités à deux intentions : énoncés initiant une prescription médicamenteuse et ceux en dehors du domaine. Afin de prendre en compte les énoncés qui ne sont pas issus du domaine de la prescription, nous avons ajouté

un corpus de parole représentatif du français parlé : ESLO2 (Serpollet et coll., 2007). Ce corpus comprend l'enregistrement et la transcription des conversations de parole spontanée produite en français. Pour que les énoncés qui représentent le hors-domaine ne soit pas plus important que celui du domaine, nous avons limité le nombre d'énoncés issus du corpus ESLO pour qu'il soit identique à ceux des prescriptions. Le tableau 4.5 présente le nombre d'exemples et de tokens issus des trois sources qui constituent notre corpus de prescriptions médicamenteuses de référence. Nous appellerons par la suite ce corpus : le corpus de **prescriptions**. Ce corpus de prescriptions est donc composé de GuideCorpus, des données artificielles générées par la grammaire et d'une partie du corpus ESLO. Pour l'apprentissage les données ont été partitionnées en ensembles de *train* (82% du total), de *dev* (4%) et de *test* (14%). Les ensembles *dev* et *test* sont composées uniquement d'exemples réalistes.

Répartition du corpus de prescriptions						
Composition	**train**		**dev**		**test**	
	nombre d'expl	nb de tokens	nombre d'expl	nb de tokens	nombre d'expl	nb de tokens
GuideCorpus	417	4696	99	1084	316	3368
Données artificielles	3034	53637	-	-	-	-
ESLO	417	2926	99	712	316	1957
Total	3868	61259	198	1796	632	5325

TABLE 4.5 – Distribution final du corpus de prescriptions de référence

Dans la répartition du corpus ESLO, nous avons pris en compte le nombre d'exemples issus du GuideCorpus (417) et non la totalité (3451) des exemples sur les prescriptions médicamenteuses. Nous avons fait ce choix pour ne pas mettre plus de ce type d'énoncés dans l'ensemble d'apprentissage pour que les données artificielles puissent jouer leur rôle d'augmentation de la couverture des attributs. L'objectif était ici d'obtenir un module de NLU suffisamment performant pour amorcer le système. Les interactions utilisateurs qui seront obtenues avec ce système permettra d'obtenir des exemples plus concrets pour ajuster le système de compréhension avec des données réelles.

4.2 Modèles automatiques de compréhension de prescriptions médicamenteuses

La tâche de compréhension automatique du langage naturel (NLU) consiste à identifier les concepts pertinents pour une tâche à partir d'un énoncé. Dans notre cas, inférer que "Paracetamol 500 mg" contient le concept `inn` dont la valeur est *paracétamol* et le concept `dose-val` dont la valeur est 500. Dans l'état de l'art, cette tâche est typiquement abordée comme une tâche d'étiquetage de séquences (*sequence labelling*) ou chaque terme de l'entrée reçoit une étiquette de l'ensemble *BIO* (*Begin-slot, Inside-slot, Outside*). Une autre information qui incombe à la NLU est la reconnaissance de l'intention portée par l'énoncé. Ainsi, dans l'exemple ci-dessus, le système doit reconnaître qu'il s'agit d'une volonté d'informer le

système de commencer une prescription.

Comme beaucoup de tâches du TALN, l'état de l'art a évolué vers des approches à base de modèle profond. Cependant, au début de ces travaux de thèse en 2018 les approchent profondes n'avaient pas encore démontrés une supériorité définitive sur les modèles classiques et l'utilisation de modèles préentraînés devait encore faire ses preuves. C'est pourquoi cette section est séparée en trois parties. La sous-section 4.2.1 rapporte l'apprentissage des modèles utilisés en début de thèse, induits à partir du corpus du domaine décrit dans 4.1.3 et qui ont servi à initier les premiers travaux sur le dialogue. La sous-section 4.3 décrit des travaux prenant en compte des données non annotées. Étant donnée la grande difficulté d'accéder sans contrainte à des données médicales en Français, ces travaux ont été réalisés sur des données étasuniennes (MIMIC). Ces travaux étant plus récents, les modèles préentrainés pour l'anglais ont été considérés.

4.2.1 Modèles initiaux de compréhension automatique

Au début du travail de thèse (2018), les meilleurs modèles *slot-filling* étaient ceux qui abordaient la tâche comme un problème multitâche. En effet, étant donné que la classification d'intention et l'extraction d'attributs sont fortement corrélées, certaines approches ont conçu des modèles pour l'apprentissage joint de ces objectifs. Ainsi ces modèles peuvent simultanément étiqueter la séquence initiale de tokens dans le format BIO et classifier une intention.

Par exemple, le modèle tri-crf (Jeong et Lee, 2008) qui est une extension du modèle CRF classique a été considéré dans ces travaux de thèse. Ce modèle d'étiquetage de séquence contient un concept supplémentaire z permettant de créer un lien de dépendance entre cette variable aléatoire z représentant l'intention et la séquence de concepts cachés qui représente les étiquettes d'attribut. Ainsi, les intentions et les attributs peuvent être appris simultanément. Bien que ce modèle soit assez ancien, ces approches à base de CRF et de règles expertes étaient encore compétitives sur le défi d'I2B2 (Uzuner et coll., 2010b) sur l'extraction d'information sur les médicaments au début de cette thèse. Les modèles CRF et tri-crf sont détaillés dans 2.5.1 de l'état de l'art.

Un autre exemple est le système neuronal multi-tâche *att-rnn* proposé par Liu et Lane (2016) qui utilise une architecture encodeur-décodeur. Ce système aborde le problème avec un encodeur unique qui alimente, d'une part, un décodeur pour étiqueter la séquence initiale dans le format BIO et, d'autre part, un classificateur d'intention. L'encodeur du modèle att-rnn est un LSTM bi-directionnel. Le décodeur est un LSTM avec attention. L'intention est déterminée par un MLP classique. Cette architecture permet un apprentissage joint de la tâche de classification d'intention et d'étiquetage des concepts qui prend en entrée les mots de l'énoncé encodant dans chaque cycle de temps un mot x_t. La sortie t de chaque cycle de temps correspond à l'état caché h_t du RNN bidirectionnel constitué de la concaténation de l'état caché de la propagation avant et arrière du RNN. Ainsi, l'état caché obtenu dans le dernier cycle de temps contient l'information sur toute la séquence d'entrée.

Enfin, nous avons considéré le modèle seq2seq (Mishakova et coll., 2019b) qui aborde le problème de NLU non pas avec un objectif d'étiquetage de séquence, mais par génération de la sémantique. Ainsi, les mots de l'énoncé sont encodés par un LSTM bi-directionnel tandis que le décodeur (un LSTM avec attention) génère la séquence d'attribut-valeur qui correspond à l'entrée. Dans ce modèle, l'intention est ajoutée à la prédiction comme un attribut en début de la séquence de sortie comme dans l'exemple ci-dessous :

```
intent [ prescription ] , inn [ paracetamol ] , d-dos-val [ 500 ] ,
d-dos-up [ milligrame ] , dos-val [ 2 ] , dos-uf [ comprimés ]...
```

Les modèles seq2seq et att-rnn sont détaillés dans 2.5.2. Dans le modèle seq2seq, contrairement à l'approche d'étiquetage de séquence de type BIO, ce type de modèle permet de prendre en compte des annotations **non-alignées**. En effet, les annotations n'ont pas besoin d'être associées à des tokens de la séquence d'origine. Ceci comporte plusieurs avantages. Premièrement, l'effort d'annotation est beaucoup moins grand que pour l'approche BIO ce qui permet d'envisager un accès à un plus vaste ensemble d'entraînement pour un coût d'annotation réduit. Deuxièmement, la génération des attribut-valeurs est indépendante de l'ordre d'entrée. Troisièmement, l'abstraction permet de générer directement des valeurs normalisées. Enfin, cet étiquetage permet au modèle seq2seq de prédire l'intention, les étiquettes et les valeurs conjointement avec un seul modèle. Sur ce dernier point les modèles tri-crf et att-rnn prédisent simultanément l'intention et les étiquettes mais pas les valeurs normalisées. Il faut ainsi apprendre deux modèles séparés pour réaliser ce triple étiquetage (intention, attributs, valeurs normalisées). L'inconvénient des systèmes non alignés est qu'ils produisent généralement des résultats de moins bonne qualité. Cependant, nous les avons considérés pour leur potentialité d'utiliser des corpus moins finement étiquetés de plus grande taille.

4.2.2 Expérience et résultats

Nous avons entraîné ces modèles sur le corpus de prescriptions décrit dans la section 4.1.3.

Pour le modèle tri-crf, afin de réduire le temps d'entraînement, nous avons éliminé les intentions ayant une probabilité inférieur à ($< 0,1\%$) et initialisé les poids en utilisant la pseudo-vraisemblance (pour 30 itérations). L'entraînement total a continué jusqu'à 200 itérations.

Dans notre implémentation de l'att-rnn, les mots d'entrée sont d'abord encodés à une couche de plongement de 128 unités. L'encodeur et le décodeur LSTM bidirectionnels constituent chacun une couche unique de 128 unités. L'entraînement est effectué par l'algorithme du gradient stochastique (SGD) avec une taille de lot de 16, en utilisant le "gradient clipping" et un *dropout* de 5.0. L'entraînement du modèle att-rnn a continué pendant 30 000 cycles d'entraînement. Puis, nous avons sélectionné le modèle qui donne le meilleur F-mesure sur les données de validation. Pour tri-crf et att-rnn, deux modèles distincts sont entraînés, un pour prédire les étiquettes, l'autre pour les valeurs portées par les étiquettes.

Pour le modèle seq2seq nous avons utilisé un encodeur biLSTM à une seule couche et un décodeur LSTM avec des plongements lexicaux de taille 128. L'entraînement a été effectué en utilisant SGD avec une taille de lot de 16. Le *learning rate* a été fixé à 0,0001 avec un *dropout* de 0.5. L'entraînement a continué jusqu'à 10000 itérations.

Enfin, pour situer les résultats par rapport à des approches classiques, nous avons considéré un modèle de CRF classique. Pour l'entraînement de ce modèle, nous avons utilisé l'outil RASA CRF qui utilise 'spacy_sklearn' pour la chaîne linéaire CRF qui classifie les attributs. Les valeurs des *slots* sont déterminées avec une table associative mise à jour automatiquement durant l'apprentissage. Pour l'intention, le modèle utilise un classifier SVM. Nous avons gardé les paramètres par défaut du modèle CRF de RASA qui s'appuie sur l'état de l'art des différents *benchmarks* (compétitions de référence).

Par ailleurs, pour déterminer l'intérêt des approches par apprentissage automatique, nous avons développé un automate à état fini, mis en œuvre par des expressions régulières. Les règles sont basées sur les valeurs fournies par la base de données médicamenteuses Thésorimed© et les exemples fréquents de rythme, de fréquence, etc. Ce modèle sera appelé par la suite modèle *baseline*.

Le tableau 4.6 présente le score de la précision, du rappel et de la f-mesure de chaque système sur le corpus de référence. Concernant la tâche de la détection d'intention, il s'agit d'une simple classification binaire. Pour la tâche de saisie d'attributs, les systèmes produisent une séquence de prédiction qui est utilisée pour mesurer la différence entre un résultat attendu et un résultat obtenu. Dans ce contexte, le résultat obtenu est la prédiction et le résultat attendu est la représentation sémantique de référence attendue.

Modèle	**Intention** (accuracy)	**Macro Moyenne**			**Moyenne Pondérée**		
		P	R	F1	P	R	F1
Baseline	-	0.33	0.29	0.28	0.54	0.47	0.49
CRF	0.97	**0.70**	0.65	0.64	**0.91**	**0.89**	0.89
Tri-CRF	0.97	0.68	**0.67**	**0.65**	0.90	0.87	0.89
Att-rnn	**0.99**	0.60	0.64	0.61	0.89	0.92	**0.91**

TABLE 4.6 – Tableau récapitulatif des performances des modèles NLU sur les attributs étiquettes du corpus de prescriptions (P=précision, R=rappel, F1=F-mesure)

Les métriques d'évaluation utilisées pour mesurer la performance des systèmes de NLU ont été détaillées dans la section 3.6.2. Sur les résultats présentés dans 4.6, nous avons ajouté un score pondéré rapport à la répartition des classes. Le score pondéré est calculé pondérant les f-mesures de chaque classe avec leurs fréquences respectives comme dans l'équation suivante :

$$F1_{classe1} * W_1 + F1_{classe2} * W_2 + ... + F1_{classeN} * W_n \tag{4.1}$$

Vu qu'il n'y a pas de distribution uniforme dans le corpus de test (il y a des attributs manquants ou très peu représentés), le Tableau 4.6 comprend aussi les scores pondérés par la fréquence des attributs.

Concernant les scores d'évaluation de la tâche de la détection d'intention, les résultats de tous les systèmes sont élevés. Cela n'est pas surprenant vu que la tâche est une discrimina-

tion binaire entre les intentions de prescriptions médicamenteuses et l'intention "none" qui ne correspond à aucune prescription. Concernant la tâche de la saisie d'attributs, les modèles entraînés avec des modèles alignés ont produit de meilleurs résultats que le modèle seq2seq. Sur les attributs-étiquettes, les modèles donnent des résultats similaires.

Les résultats sur le Tableau 4.6 montre qu'au niveau de la micro-moyenne, le meilleur système de NLU est att-rnn. Ces résultats montrent également qu'avec un nombre limité de données, le système att-rnn réussit à apprendre des modèles performants. Cela est confirmé par le score de la moyenne pondérée qui favorise les classes qui sont représentées plus que d'autres.

Le modèle seq2seq est celui qui est plus proche par rapport au *baseline* à base de règles. Cela contredit les résultats récents obtenus par un système similaire Desot et coll. (2018), cependant cela pourrait être lié à la taille du corpus d'entraînement qui est relativement petite pour l'apprentissage des modèles non-alignés étant donné que dans ce cas, l'alignement doit également être appris par le modèle. En revanche, d'un point de vue macro, les systèmes classiques à base de CRF sont plus performants que att-rnn. Les attributs qui sont peu représentés sont mieux reconnus par ces systèmes.

Modèle	**Micro Moyenne**			**Macro Moyenne**			**Moyenne Pondéré**		
	P	R	F1	P	R	F1	P	R	F1
CRF	0.70	0.68	0.69	0.29	0.29	0.29	0.85	0.68	0.73
Tri-CRF	0.86	0.82	0.84	0.39	0.40	**0.38**	0.84	0.82	0.83
Att-rnn	0.85	0.87	**0.86**	0.32	0.36	0.32	0.85	0.87	**0.85**
Seq2Seq	0.58	0.57	0.57	0.14	0.13	0.13	0.58	0.57	0.57

TABLE 4.7 – Tableau récapitulatif des performances des modèles NLU sur les attributs valeurs du corpus de prescriptions

Le Tableau 4.7 montre les performances de ces systèmes sur les attributs-valeurs normalisés. L'ordre de performance des systèmes reste le même. Cependant, la prédiction des valeurs normalisées étant plus difficile que les étiquettes, les scores sont plus faibles pour tous les systèmes. Nous remarquons également que pour les systèmes classiques, tri-crf est beaucoup plus performant sur les valeurs normalisées qu'un simple modèle CRF. De la même manière, au niveau macro tri-crf semble donner de meilleurs résultats comparés au modèle att-rnn(6% plus performant).

La Figure 4.3 montre la matrice de confusion issu des prédictions du modèle tri-crf.

Malgré la distribution non uniforme des classes de notre corpus de test, la Figure 4.3 nous montre que le modèle prédit les étiquettes des attributs avec une bonne performance. Dans ce tableau, les vrais étiquettes sont représentés à gauche (sur l'axe vertical) et les étiquettes prédits sont en bas (sur l'axe horizontal). Comme dans une tâche de NER classique, nous représentons explicitement les mots hors vocabulaire "**O**=Outside" est l'attribut qui comporte le plus d'exemples. Cette sur-représentation est partiellement lié à l'inclusion du corpus ESLO2 qui représente 50% du corpus de test. En revanche, contrairement aux comptes rendus médicaux, les prescriptions sont condensés en termes d'information sémantique. Le tableau 4.8 montre la répartition détaillé du corpus de test.

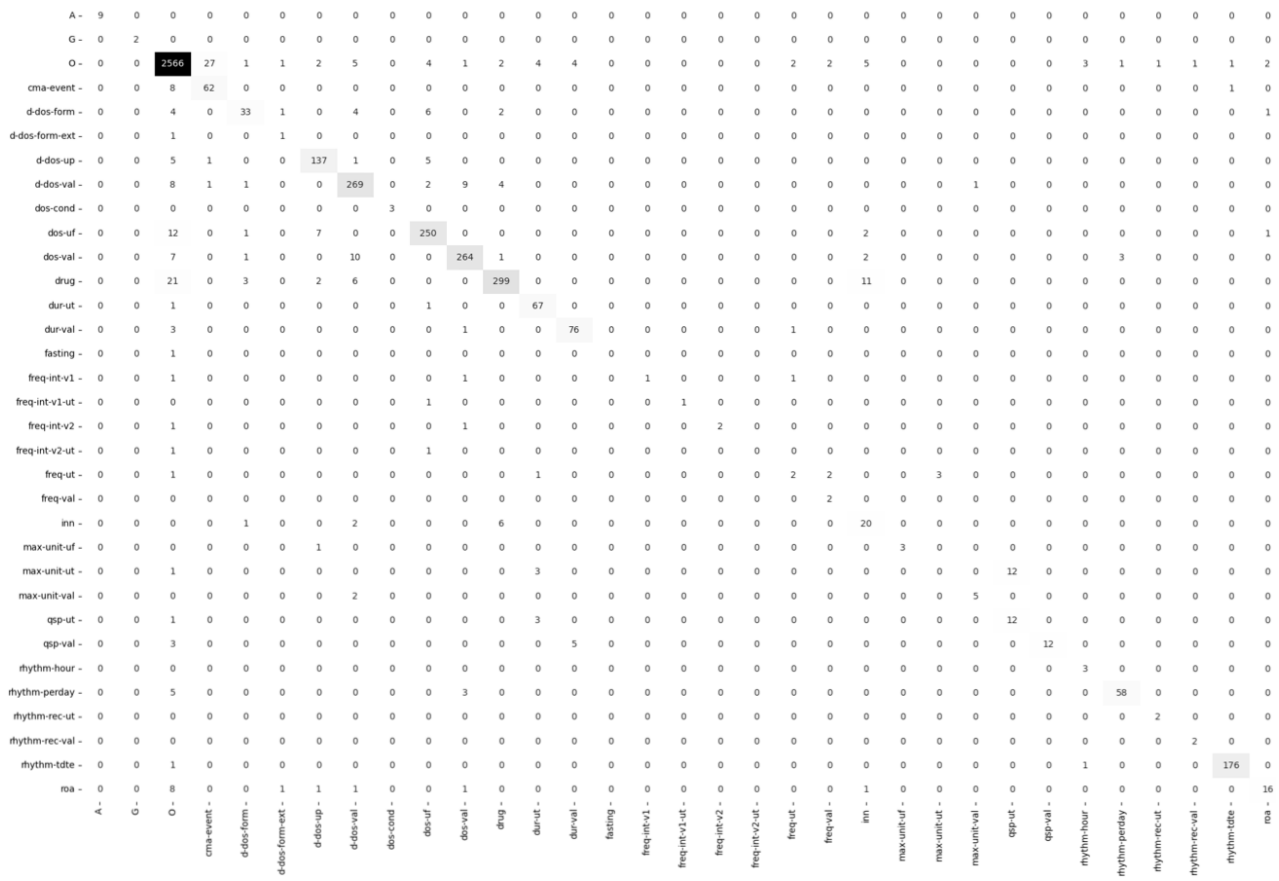


FIGURE 4.3 – Matrice de confusion des prédictions du modèle tri-crf

Nombre d'attributs-étiquettes	1783 attributs
Nombre d'attributs max/prescription	17 attributs
Nombre d'attributs min/prescription	1 attributs
Moyen du nombre d'attributs/prescription	0.62 attributs/tokens
Longueur moyen des prescriptions	9.62 tokens

TABLE 4.8 – Distribution de nombre d'attributs par rapport au tokens parmi les prescriptions du corpus de test

La synthèse présentée sur le Tableau 4.8 montre que plus que la moitié des mots dans une prescription correspond à un attribut. Le nombre total d'attributs (1783) sur 316 prescriptions montre que les prescriptions véhiculent beaucoup d'information sémantique concernant le médicament, la posologie et les informations liées à l'administration. On peut conclure que le nombre de confusions montrées dans le Tableau 4.3 est relativement bas par rapport au nombre total d'attributs. En revanche, parmi les attributs qui sont confondus, nous pouvons remarquer que l'attribut "drug" (nom de spécialité) est plus souvent confondu avec un DCI (inn) ou est considéré comme un mot hors vocabulaire. De la même manière, la matrice de confusion montre que la forme de la posologie (dos-uf) est plus souvent confondue que la forme (d-dos-form) du médicament. Cela est peut-être lié aux formulations de la posologie qui comportent moins de contraintes comparées aux formes galéniques. Par exemple, « paracétamol 1 comprimé, 2 **cachets** le matin ... » Cela n'est pas étonnant, vu que la posologie est destinée au patient, elle est souvent exprimée en une langue plus naturelle.

L'autre catégorie de confusion est liée aux valeurs des unités de temps(*-val). Sur une prescription, il peut y avoir plusieurs indications liées à la fréquence, le rythme et la durée qui comportent une sémantique relativement proche qui résulte dans des confusions.

Même si les systèmes de NLU comparés dans cette section donnent des résultats globalement très proches, les analyses que nous avons faites nous montrent que les approches tri-crf et att-rnn réussissent mieux à généraliser sur les données de test du corpus de prescriptions. Parmi ces deux approches, att-rnn semble être plus performant à une échelle de micro moyenne ou lorsque les classes sont pondérées. En revanche, à l'échelle de macro moyenne, le modèle tri-crf semble être plus performant pour les attributs étiquettes et les valeurs normalisées. Comme nous l'avons évoqué précédemment, pour la validation d'une prescription, tous les attributs sont importants, même ceux qui sont rares. C'est pourquoi, dans la mise en place initiale du système de dialogue, nous avons choisi d'utiliser le modèle tri-crf.

4.3 Modèles de compréhension automatique par apprentissage semi-supervisée

Comme dit à plusieurs reprises dans ce manuscrit, l'accès à des données médicales annotées en Français est très difficile. Cependant, dans le cadre de cette thèse CIFRE, l'entreprise a accès à un ensemble de données propriétaires non annotées qui pourraient être exploité pour apprendre des modèles de NLU plus robustes. Cette situation est en fait bien plus courante et réaliste que le cas standard de l'apprentissage supervisé où un grand ensemble de données annotées est disponible.

Cependant, afin de préalablement tester cette idée, nous avons développé plusieurs approches permettant d'utiliser un petit ensemble de données annotées et un grand ensemble de données non annotées. Pour cela, nous avons utilisé des corpus en Anglais proches de notre cas d'étude, celui du défi d'I2B2-2009 et MIMIC-III présentés dans 2.6. Concernant le choix d'un corpus annoté, nous avons choisi le corpus I2B2 qui est annoté en information sémantique sur le médicament. Même si formellement il ne s'agit pas d'un corpus de prescriptions médicamenteuses, les textes du corpus contiennent des prescriptions ou des informations sur la prise de médicaments. I2B2-2009 est considéré comme un *benchmark* pour la tâche d'extraction d'information médicamenteuse et sera notre corpus de test et de données d'entraînement alignées. Pour les données non annotées, l'une des ressources plus riches utilisées dans la plupart des domaines de l'informatique biomédicale est la base de données MIMIC-III. Cependant, ce corpus est purement textuel et donc non adapté à notre cas d'énoncés oraux. La sous-section 4.3.1 présente comment nous avons préparé ces corpus pour qu'ils soient plus proches de nos cas d'usage oral.

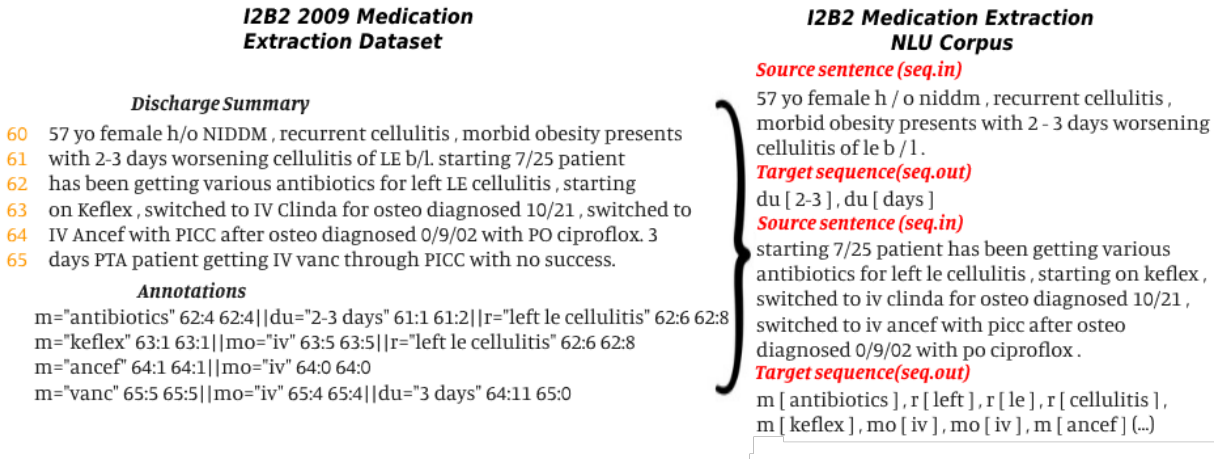
Pour tirer parti de données non annotées, nous avons utilisé deux approches : une approche semi-supervisée et une approche par *fine tuning* de modèles pré-entrainés. La méthode semi-supervisée est décrite en sous-section 4.3.2, elle a été proposée par Qader et coll.

(2019) et consiste en un couplage d'un système de NLU et de NLG (*Natural Language Generation*). Les modèles pré-entrainés sont présentés en sous-section 4.3.3. Les résultats de l'ensemble de ces modèles sont présentés et discutés en sous-section 4.3.4.

4.3.1 Préparation des corpus

Dans notre cas de compréhension automatique des prescriptions orales, les entrées sont des énoncés oraux. Cependant, les corpus I2B2 et MIMIC-III sont composés de comptes rendus médicaux (EHR) qui sont distribués sous forme de documents textuels où chaque annotation indique les positions exactes des segments dans le texte concerné par l'annotation (index ligne : index mot).

Nous avons appliqué une méthode permettant d'extraire les segments textuels de prescription des documents en énoncés isolés du contexte. Cette méthode permet également de convertir les annotations des documents en annotation de type non alignée pour le format séquence à séquence. La Figure 4.4 montre un exemple d'annotation sémantique d'un document I2B2 et sa conversion au format seq2seq proposé par Kocabiyikoglu et coll. (2021).



À gauche : un compte rendu issu du défi d'extraction de médicaments d'I2B2-2009 et les annotations sémantiques indexées. À droite : Les mêmes segments représentés par des séquences d'entrée et de sortie adaptée à un format seq2seq. (m = nom du médicament; do = dosage; mo = mode d'administration; f = fréquence; du = durée; r= raison)

FIGURE 4.4 – Résumé du processus de préparation de données alignées au format séquences à séquences (Kocabiyikoglu et coll., 2021)

La Figure 4.4 nous montre un exemple décrivant le processus de transformation d'une phrase comportant une annotation indexée par (ligne : mots) vers un format seq2seq. Cet exemple montre bien que dans un compte rendu, il y a beaucoup de texte en dehors de la sémantique qui n'est pas annoté. C'est-à-dire qu'un segment source relativement long peut-être associé à une séquence de sortie plus courte comme illustré sur cet exemple issu du défi d'I2b2-2009.

Les éléments considérés dans le défi I2B2-2009 (Uzuner et coll., 2010b) sont les *slots* suivants : **(m)** : médicament (spécialité, DCI, générique, produit en accès libre, etc.), **(do)** : dosage du médicament ou de la posologie, **(mo)** : la voie d'administration, **(f)** : la fréquence de la prescription, **(du)** : durée de la prise du médicament et **(r)** : la raison pour laquelle le médicament est donné.

Le tableau 4.9 présente les caractéristiques des données qui ont été distribuées et montrent qu'une partie importante des données a été mise de côté par les organisateurs du défi (Uzuner et coll., 2010b,a). En fait, dans ce défi, le corpus de test de référence est celui annoté par la communauté. À la fin de la compétition, une partie des comptes rendus ont été mis de côté pour être annotés par la communauté mais est resté non utilisé. Ces données-là sont indiquées comme données supplémentaires dans le Tableau 4.9.

Distribution des données	I2B2-2019				MIMIC-III
	Supervisé			Semi-Supervisé	Semi-Supervisé
	annotations de la communauté	annotations expertes	données non annotés	données supp. non-annotées	notes cliniques
train	-	-	22,907 phrases	4,655 phrases	257,811 phrases
validation		148 phrases	2,158 phrases	-	2,605 phrases
test	4411 phrases			-	-

TABLE 4.9 – Répartition des corpus I2B2-2009 et MIMIC-III et distribution dans l'ensemble supervisé/non-supervisé de notre étude.

Comme l'approche seq2seq fonctionne au niveau de la phrase, nous avons effectué une extraction au niveau des phrases des documents bruts des deux corpus (I2B2-2009 & MIMIC) à l'aide de la boîte à outils *ClarityNLP*[1] qui s'appuie sur *Spacy* que nous avons initialisés avec un modèle de langue BERT (en_core_web_trf). Sur les phrases extraites, nous avons appliqué les annotations qui sont au format (ligne :mots) à un format seq2seq comme sur la Figure 4.4. Enfin, pour traiter les mots hors vocabulaire (OOV), nous avons utilisé un codage par paire d'octets - *byte-pair encodings* (BPE) (Sennrich et coll., 2015). Les codages ont été appris à partir des corpus de textes MIMIC-III et I2B2 (ensemble de test exclu). En effet, même si l'ensemble de données d'I2B2 est petit, la taille du vocabulaire et les mots rares de la séquence d'entrée sont élevés (Zhang et coll., 2019b).

Même si MIMIC-III (Johnson et coll., 2016) est une source de données cliniques riche, il ne contient pas de données adaptées à l'apprentissage automatique supervisé vu qu'il ne contient pas de données annotées. En revanche, il contient des informations sur les médicaments à la fois dans la partie textuelle des comptes rendus cliniques et dans la partie base de données (table des médicaments prescrits). Concernant les informations sémantiques sur les médicaments, les données structurées fournissent les informations suivantes : nom du médicament, la date du début et de la fin de l'administration sous forme de *timestamp*, nom du générique, dosage du médicament et dose, route, forme de la posologie. En fait, ces données-là sont très proches des données que nous pourrons trouver dans un LAP.

Pour l'apprentissage semi-supervisé, il a fallu effectuer un prétraitement pour extraire les informations pertinentes liées aux prescriptions de médicaments. Pour ce faire, nous avons mis en place un algorithme de positionnement (ranking) simple : pour chaque prescription dans la base de données, les lignes du compte rendu du même patient ont été notées à l'aide d'expressions régulières. La Figure 4.5 illustre cette stratégie d'association.

Prenons l'exemple de la Figure 4.5. La partie droite montre un extrait de la table des prescriptions, tandis que la partie gauche montre quelques phrases du résumé de sortie du pa-

1. https://claritynlp.readthedocs.io/en/latest/

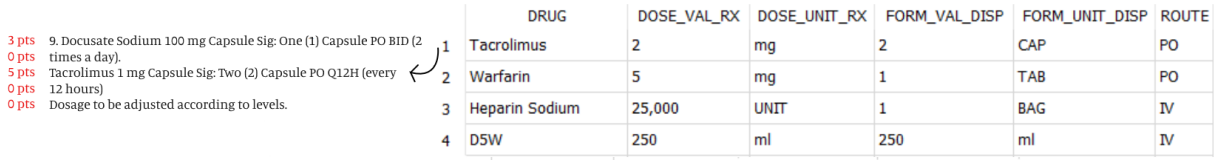


FIGURE 4.5 – Exemple d'association des comptes rendus médicaux aux enregistrements de la base de données lié aux médicaments

tient qui ont été évaluées. La phrase ayant obtenu 5 points était celle qui correspondait le plus aux caractéristiques de la phrase Tacrolimus. Ainsi, la phrase et la ligne de prescription sont ajoutées à l'ensemble de données annotées. Les prescriptions qui ne correspondent à aucune ligne du résumé de sortie ne sont pas ajoutées à l'ensemble de données non annotées.

Au niveau sémantique, les informations enregistrées dans la base de données MIMIC-III ne sont pas les mêmes que celles du jeu de données d'extraction de médicaments d'I2B2. Par exemple, dans la Figure 4.5, le texte comprend des informations sur la fréquence du traitement, mais la base de données ne comporte aucune colonne pour décrire cette entité. De même, nous ne disposons que d'informations systématiques normalisées par les normes RxNorms (Liu et coll., 2005) couvrant le médicament, la posologie, le mode d'administration et la durée du traitement. Cependant, l'information sur la durée est dans un format de date et n'est pas donnée en langage naturel, elle a donc été exclue des données. En outre, les descriptions sémantiques sur les fréquences et les raisons de la médication étaient absentes des données.

À la fin du processus, 962 252 lignes de phrases ont été extraites à partir de la base de données. Elles ont ensuite été filtrées afin de supprimer les exemples similaires ainsi que les énoncés trop longs. Certaines phrases transmettent beaucoup plus d'informations que celle de la ligne équivalente, ce qui rend les données bruitées et répétitives. Ainsi, nous avons réduit la taille du corpus en excluant les paires phrase-sémantique strictement équivalentes et pour obtenir un corpus final de 260 416 phrases et lignes de base de données faiblement couplées.

Il convient de noter que les informations sémantiques de la base de données MIMIC-III sont différentes de celles d'I2B2-2009. Par exemple, il n'y a pas de raison ou de fréquence dans les enregistrements du MIMIC-III. Cependant, ce type d'information est souvent présente dans la partie textuelle de MIMIC. On peut donc en conclure que ce corpus est pertinent pour la tâche d'extraction de médicaments.

Dans notre approche, afin de constituer la partie non alignée du corpus semi-supervisé, nous avons mis en place une technique simple de positionnement de phrases alimentée par les expressions régulières. Pour ce faire, nous avons d'abord extrait les comptes rendus d'hospitalisation des patients avec la liste de médicaments prescrits. Les comptes rendus ont été tokénisés en phrases en utilisant la librairie *Clarity NLP* qui s'appuie sur *Spacy* que nous avons initialisés avec un modèle de langue BERT.

4.3.2 Approches semi-supervisées

Concernant les approches semi-supervisés, Tao et coll. (2018) ont proposé un système semi-supervisé qui a atteint la meilleure performance globale actuelle sur la tâche d'extraction de médicaments d'I2B2-2009 en exploitant la partie non annotée du corpus. Plus récemment, Guzman et coll. (2020) ont proposé un système basé sur un modèle LSTM et l'apprentissage par transfert qui affirme des performances état de l'art sur l'extraction d'entités spécifiques. Cependant, leur système n'est que partiellement décrit et donc difficile à reproduire. Pour d'autres tâches, telles que l'extraction de relations biomédicales, des autoencodeurs variationnels (Zhang et Lu, 2019) ou des modèles EFCG (*Event Feature Coupling Generalization*) ont été proposés pour tirer profit des données non annotées. En particulier, Amin et coll. (2020) montrent que pour la tâche d'extraction de relations biomédicales, une approche supervisée à distance permet de produire de grandes quantités de données étiquetées mais bruitées qui peuvent être exploitées efficacement pour une approche guidée par les données. Malgré ces progrès récents, l'apprentissage semi-supervisé pour l'extraction de médicaments n'a été appliqué que par Tao et coll. (2018), ce qui représente à ce jour l'état de l'art.

Dans ce travail de thèse, nous avons utilisé l'approche semi-supervisée seq2seq proposée par Qader et coll. (2019). Cette approche considère deux modèles distincts d'encodeur-décodeur : l'un pour extraire la sémantique à partir du texte (NLU) et l'autre pour générer du texte à partir d'une représentation de sens (NLG). L'approche considère trois ensembles de données : du texte aligné avec les annotations (I2B2-2009), du texte non aligné (comptes rendus de MIMIC-III) et des annotations sémantiques non alignées (la table des médicaments prescrits aux patients du MIMIC-III). Les données alignées sont utilisées pour apprendre de manière supervisée les modèles de NLU et de NLG. Les données non annotées sont utilisées par les deux modules de NLU et de NLG pour générer leurs contre-parties respectives. Le texte (resp. la sémantique) en entrée est envoyé aux modèles NLU (resp. NLG) qui produisent une représentation sémantique (resp. un texte) qui est à son tour envoyé au NLG (resp. NLU) qui produit un texte (resp. une représentation sémantique). La différence entre les textes d'entrée et de sortie (resp. sémantique) est utilisée comme une fonction de coût (loss) pour optimiser les deux modules conjointement. De cette façon, les données qui ne sont pas alignées à une annotation peuvent être utilisées pour l'apprentissage en utilisant cet objectif de « reconstruction ». Cette stratégie de l'apprentissage jointe de NLU et de NLG est schématisée sur la Figure 4.6.

La Figure 4.6 décrit cette méthode qui consiste à produire du texte $\hat{y}$ à partir d'une représentation de sens x qui est passé au module de NLU qui produit une représentation de sens $\hat{x}$ à partir du texte y en utilisant 2 encodeurs décodeurs en parallèle. Comme les modèles de NLU et de NLG sont appris conjointement, les *loss* des modèles NLG et NLU pour les modèles alignés et non alignés pourraient être notées respectivement comme L_p^{nlg},L_p^{nlu}, L_u^{nlg} et L_u^{nlu}(p= données alignées, u = données non alignées). Ces quatre *loss* sont mis ensemble pour effectuer l'apprentissage joint $L = \alpha$ en concaténant de façon pondérée ces différents *loss* de

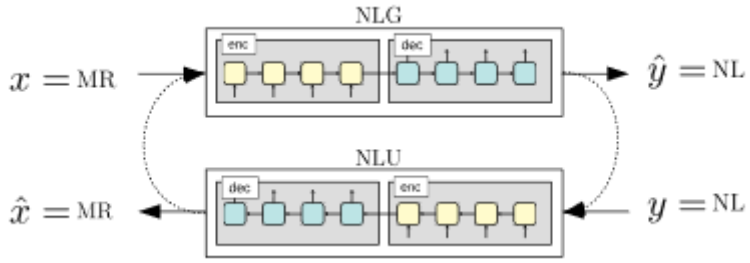


FIGURE 4.6 – Apprentissage joint de NLU et de NLG proposé par Qader et coll. (2019)

comme suivant :

$$L = \alpha L_p^{nlg} + \beta L_p^{nlu} + \gamma L_u^{nlg} + \delta L_u^{nlu} \tag{4.2}$$

Les poids α, β, γ et $\delta \in [0,1]$ sont définis pour effectuer un *finetuning* déterminant ainsi la contribution des tâches spécifiques au processus d'apprentissage conjoint.

4.3.3 Approches à base de modèles de langues pré-entraînés

Récemment, les modèles de langues spécialisées dans le domaine biomédical ont donné des performances très prometteuses dans les différents défis d'I2B2 (Jin et coll., 2019; Lee et coll., 2020; Peng et coll., 2019). Sur la tâche d'extraction de médicaments d'I2B2, Tao et coll. (2017) évalue les performances des différents classifieurs tels que le modèle bayésien naïf multinomial, les SVM, les arbres de décision, et les CRF. Dans cette approche, ils entraînent des plongements lexicaux en se basant sur le corpus de MIMIC-III pour apprendre des représentations vectorielles avec *Glove*. Comme dans Tao et coll. (2018), l'apprentissage s'appuie sur sous un ensemble très petit du corpus d'entraînement annoté par des experts médicaux issu de Patrick et Li (2009). De façon similaire, Gligic et coll. (2020) propose une approche à base de RNN avec des plongements *word2vec* entrainés sur la partie du corpus non utilisé d'I2b2-2009.

Ces approches, même récentes s'appuient sur des méthodes plus classiques comme les SVM et les RNN. Pour avoir un point de comparaison et découvrir les performances des modèles de langues à base de transformeur, nous avons effectué un *fine-tuning* sur la tâche d'extraction d'information sur le médicament. Afin d'établir une référence, le premier modèle que nous avons inclut est un modèle de langue d'usage général : BERT (Vaswani et coll., 2017). Même si BERT n'est pas entrainé sur un corpus médical, il permettrait justement de comparer l'avantage des modèles de langues spécialisés dans le domaine médical. Comme le modèle de langue biomédicale, nous avons inclut BioBert (Lee et coll., 2020), BlueBert (Peng et coll., 2019) et Clinical Bert (Alsentzer et coll., 2019).

Récemment, Zhu et coll. (2020) ont proposé une nouvelle approche pour la traduction automatique neuronale dans lequel ils exploitent les plongements de BERT, d'abord par récupérant des représentations à partir d'une séquence d'entrée, puis en les fusionnant avec chaque couche de l'encodeur et du décodeur par le biais d'un mécanisme d'attention après. Nous avons considéré que ce modèle peut établir un point de comparaison plus proche à

seq2seq en s'appuyant justement sur un modèle de langue entraîné sur l'architecture transformeur. Cette approche de fusion est schématisée sur la Figure 4.7.

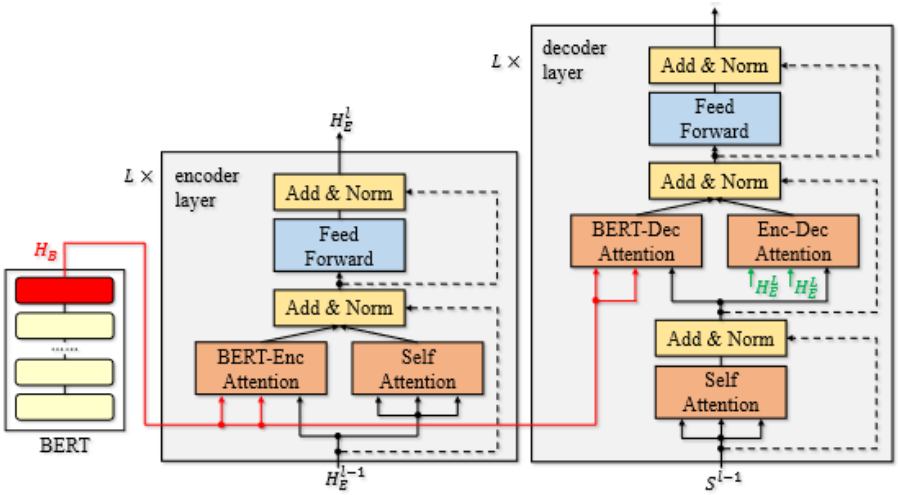


FIGURE 4.7 – Principe de fonctionnement du modèle de fusion de BERT dans un cadre séquence à séquence (Zhu et coll., 2020)

Sur la Figure 4.7, la partie à gauche représente l'encodeur et la partie à droite représente le décodeur. Dans cette approche, tout entrée $x \in X$ est traité progressivement par l'encodeur et décodeur BERT. Sur la Figure 4.7, les lignes pointillées dénotent les connexions résiduelles. H_b (la partie rouge) et H_E^L (la partie verte du décodeur) dénotent la sortie de la dernière couche de l'encodeur BERT.

4.3.4 Résultats des modèles pré-entrainés et semi-supervisés sur le corpus I2B2-2009

Pour la tâche d'extraction d'entités nommées autour de médicament, nous avons évalué des modèles seq2seq (non alignée), la méthode semi-supervisée présentée dans 4.3 et les modèles de langues présentés dans 2.8.2 *finetuné* sur le corpus d'I2B2-2009.

Concernant les méthodes supervisées, la première méthode que nous avons incluse est un modèle classique de seq2seq bi-directionnel avec du mécanisme d'attention proposé par Luong et coll. (2015). Ce modèle est capable d'apprendre des dépendances en court et long terme efficacement et peut être entraîné sur une quantité raisonnable de données. Il est également étonnamment efficace. Nous avons également inclus un modèle CNN, car les CNN sont capables de capturer les relations hiérarchiques entre les entités et sont assez efficaces à entraîner. Nous avons également considéré un modèle convolutionnel basé sur une architecture seq2seq (conv-s2s) de Gehring et coll. (2017). Enfin, nous avons inclus un modèle basé sur l'architecture transformeur Vaswani et coll. (2017) qui fait partie des modèles actuels les plus performants sur plusieurs tâches. Il est à noter que sur cette même tâche de *NER* sur l'extraction de médicaments, les modèles seq2seq sont entrainés de façon non alignée alors que les autres modèles sont entraînés de façon alignés. L'entraînement est effectué en utilisant deux bibliothèques de seq2seq libres de droits : la bibliothèque seq2seq-py

de Qader et coll. (2020) et la bibliothèque fairseq de Ott et coll. (2019). Pour le *finetuning* des modèles de langues pre-entraînés sur le corpus d'I2B2, nous avons utilisé la librairie de simpletransformers [2].

La comparaison des performances de ces approches supervisées classiques, et les modèles utilisant des modèles de langues générales et spécialisés dans le domaine biomédical sont présentés sur le Tableau 4.10.

Modèle	**F1**	**m**	**do**	**mo**	**f**	**du**	**r**
LSTM*	0.78	**0.94**	**0.92**	0.93	0.89	**0.49**	0.50
LSTM(bpe)*	0.75	0.88	0.90	0.91	0.88	0.46	0.46
conv-s2s (bpe)†	0.68	0.87	0.83	0.84	0.76	0.38	0.41
transformer (bpe)†	0.75	0.92	0.88	0.89	0.84	0.47	0.50
Modèles Pre-entrainés	**F1**	**m**	**do**	**mo**	**f**	**du**	**r**
bert-base†	0.63	0.85	0.85	0.82	0.83	0.28	0.17
bert-fused -transformer†	0.74	0.90	0.87	0.89	0.83	0.47	0.50
clinical-bert†† base	0.75	0.82	0.76	0.75	0.76	0.33	0.45
biobert-base††	0.75	0.82	0.76	0.75	0.76	0.30	0.44
bluebert-†† base	**0.88**	0.92	0.88	**0.95**	**0.91**	0.46	**0.61**

* :Librairie seq2seq-py, † : Librairie fairseq, †† Librairie Simpletransformers
(m=médicament, do=dosage, mo=voie d'administration, f=fréquence, du=durée, r=raison)

TABLE 4.10 – Performances (F-mesure) des modèles de la tâche d'extraction d'information sur le médicament d'I2B2-2009

Comme présenté sur le Tableau 4.10, les résultats présentent le F-mesure sur la globalité des phrases ainsi que pour chacune des entités nommées à extraire. Comparées au domaine sémantique sémantique des prescriptions médicamenteuses, dans cette tâche, les informations sont présentées de façon plus globale. Par exemple, tout médicament, produit thérapeutique ou DCI peut rentrer dans le cadre du *drug* alors que notre modélisation représente chacune différemment pour permettre une intégration complète à un plan de soins. La partie haute du tableau 4.10 présente les performances des modèles utilisant une approche classique supervisée : LSTM, CNN et Transformeur. Les modèles entraînés suite à l'application d'un codage par paire d'octets (bpe) réduisent la taille du vocabulaire et considèrent mieux les mots hors vocabulaires. Ces résultats nous montrent que parmi les méthodes supervisées classiques, LSTM simple avec de l'attention obtient f-mesure le plus élevé dans la globalité ainsi que sur toutes les *slots* individuellement. Cela pourrait être lié à la définition du domaine qui est assez spécifique et la quantité de données d'apprentissage qui est faible. Pour la même raison, l'utilisation de BPE n'apporte pas d'amélioration. Dans les tâches du TAL général, les méthodes CNN et transformeur sont utilisées lorsqu'il y a un ensemble de don-

2. https://github.com/ThilinaRajapakse/simpletransformers

nées plus importants mais sur cette tâche avec la faible quantité de données annotées, ils ont été moins efficaces.

Les performances des modèles de langues entraînées sur des textes de façon non supervisée, spécialisés dans le domaine général ou médical, sont présentées en bas du Tableau 4.10. Le modèle de langue d'usage général classique *bert-base-uncased* n'a pas réussi à se spécialiser suffisamment tandis que l'utilisation fusionnée du même modèle proposé par Zhu et coll. (2020) n'a atteint que des performances comparables à celles des méthodes supervisées standards. Comme pour les méthodes supervisées, le manque de données pourrait expliquer la faible performance des modèles pré-entraînés d'usage général. Cependant, le modèle (*bluebert-base-uncased*), qui a été spécifiquement pré-entraîné sur des documents médicaux (PubMed et MIMIC) a atteint la meilleure mesure F (88%) et s'est montré particulièrement performant pour les étiquettes de fréquence (f) et de raison (r), connues pour être particulièrement difficiles.

L'utilisation d'autres ressources textuelles pour améliorer les performances des tâches spécifiques a l'air de fonctionner sur cette tâche d'extraction de médicaments montrés par les résultats élevés (Peng et coll., 2019; Tao et coll., 2017). Les résultats de l'approche semi-supervisée présentée dans 4.3 sont présentés sur le Tableau 4.11.

Modèle	α	β	γ	δ	**F1**	**m**	**do**	**mo**	**f**	**du**	**r**
mimic (bpe)	1	0.1	1	0.1	0.55	0.83	0.77	0.79	0.75	0.37	0.39
mimic	1	0.1	1	0.1	0.64	**0.92**	0.88	0.90	0.85	**0.47**	**0.46**
i2b2 (bpe)	1	0.1	1	0.1	0.73	0.90	0.88	0.89	0.87	0.46	0.42
i2b2	1	0.1	1	0.1	**0.74**	0.91	**0.89**	**0.91**	**0.87**	0.43	0.44

TABLE 4.11 – F-mesure des modèles semi-supervisés sur le corpus de test d'I2B2-2009 basée sur les LSTM avec de l'attention

Les modèles semi-supervisés présentés sur le Tableau 4.11 s'appuient sur la meilleure méthode supervisée (LSTM + attention). Comme pour les autres modèles présentés, la f-mesure de l'ensemble de la tâche ainsi que des *slots* sont présentées sur le Tableau 4.11. Les valeurs des α, β, γ et δ ont été apprises empiriquement. Ces poids permettent d'orienter l'apprentissage joint (la compréhension ou la génération) plus ou moins vers l'un (NLU) ou l'autre (NLG). Les résultats montrent que lorsque l'apprentissage semi-supervisé est effectué sur le corpus MIMIC-III, les résultats globaux sont décevants. Cela est dû au fait que le corpus de MIMIC-III et I2B2 sont encore trop divergents. Par exemple, étant donné que MIMIC-III est une base de données issue d'un hôpital, beaucoup de prescriptions ont une forme très standardisée. En revanche, les prescriptions issues d'I2B2 sont souvent sous forme narrative. Étant donné que MIMIC est un ensemble de données beaucoup plus riche, l'utilisation de BPE apporte une nette amélioration (très bonne détection des noms de médicaments). Comme Gligic et coll. (2020), nous avons considéré l'utilisation de la partie non annotée du corpus d'I2B2-2009 pour apprendre des modèles semi-supervisés. Ces modèles sont présentés en bas du Tableau 4.10. Lorsque l'apprentissage semi-supervisé est réalisé sur cet ensemble de données non annotées au lieu de MIMIC, les performances ont augmenté.

Cela est dû à une bonne correspondance entre les données d'entraînement et les données de test. Cependant, la performance n'atteint pas celle de la méthode supervisée classique. Ainsi, il semble que pour cette tâche, l'utilisation d'un pré-annotateur comme MedExtractor (Doan et coll., 2009) est plus efficace que la stratégie semi-supervisée. Cependant, pour les langues pour lesquelles un tel extracteur n'existe pas, la stratégie semi-supervisée représente une bonne alternative. Le tableau 4.12 présente les performances des meilleurs systèmes en comparaison avec les meilleurs modèles que nous avons entraînés sur I2B2-2009.

Système	F1	m	do	mo	f	du	r
Guzman et coll. (2020)	0.76	0.78	0.81	0.78	0.82	0.19	-
Tao et coll. (2018)	0.87	0.93	**0.94**	**0.95**	**0.94**	**0.68**	0.48
LSTM + attention	0.78	**0.94**	0.92	0.93	0.89	0.49	0.50
bluebert-base (Peng et coll., 2019)	**0.88**	0.92	0.88	0.95	0.91	0.46	**0.61**

TABLE 4.12 – Comparaison de deux systèmes état de l'art comparé à nos 2 meilleurs modèles sur la tâche d'extraction de médicaments d'I2B2-2009

Le tableau 4.12 résume les meilleurs résultats que nous avons obtenus et les compare avec l'état de l'art actuel. Nous pouvons constater qu'un simple modèle bi-LSTM donne des résultats compétitifs. Notre modèle pré-entraîné basé sur BlueBert bat le modèle de Tao et coll. (2018) par une courte marge. Toutefois, il est important de noter que Tao et coll. (2018) a utilisé un ensemble de données d'entraînement annotées par des humains experts dans le domaine biomédical, alors que notre approche n'a utilisé que des données annotées automatiquement. En outre, le modèle BlueBert montre une grande capacité à extraire le concept de raison qui a été constaté comme le plus difficile à traiter dans le défi I2B2-2009.

4.4 Conclusion

Dans ce chapitre, nous avons présenté la tâche de compréhension de prescriptions médicamenteuses orales connectées à un plan de soins dans un contexte de faibles ressources. Nous avons détaillé comment nous avons acquis le premier jeu de données que nous avons extrait à partir d'un livre.

Ce premier jeu de données représentait très peu d'exemples en termes de quantité pour entraîner des modèles. De plus, la répartition des classes était très inégale. Or, pour un domaine aussi sensible que le domaine médical, un système devrait être capable de reconnaître parfaitement toutes les informations même celles qui sont rares. Face à ce problème, nous avons présenté notre méthode de génération de données artificielle qui a permis d'augmenter la quantité de données ainsi que la couverture des *slots* peu représentés.

Ce jeu de données nous a permis d'apprendre des premiers modèles de compréhension dans le domaine des prescriptions médicamenteuses. Nous avons entraîné plusieurs modèles de compréhension par apprentissage supervisé avec un objectif d'étiquetage aligné tel que tri-crf, att-rnn et un objectif d'étiquetage non aligné avec le modèle seq2seq. Les performances de ces modèles sur les prescriptions extraites à partir des livres pédagogiques nous

montrent que même avec un nombre limité de données, il est possible de capturer beaucoup d'information sémantique sur une prescription médicale. Malgré quelques confusions et attributs plus difficiles à reconnaître montrés en section 4.2.2, un système de dialogue pourrait permettre la correction et la complétion de ces informations afin d'établir une intégration complète avec un système hospitalier.

Enfin, pour tirer parti d'un grand nombre de données non-annotées, nous avons fait une comparaison des méthodes supervisées, semi-supervisées et par transfert (modèles de langues pré-entrainés) sur une tâche de NER qui est proche de la sémantique des prescriptions médicamenteuses : I2B2-2009 défi sur l'extraction d'information sur le médicament. Les modèles que nous avons entraînés ont montré des résultats prometteurs malgré l'utilisation d'annotations automatiques montrant qu'on peut obtenir des résultats état de l'art en utilisant le transfert de connaissances. Les résultats de l'approche semi-supervisé n'étaient pas à la hauteur, notamment comparé aux modèles de langues pré-entrainés, cependant ils représentent une alternative qui permet ce transfert de connaissance en l'absence des modèles de langue pre-entrainés, notamment pour d'autres langues que l'anglais.

Dans ce chapitre, la compréhension s'est concentrée sur la compréhension des prescriptions complètes, cependant dans un contexte dialogique, il y a d'autres intentions et d'autres interactions conversationnelles que le système de compréhension devraient extraire l'intention et la sémantique. Même pour les prescriptions, un prescripteur aura besoin de compléter ou corriger des informations. Le chapitre 5 abordera cet aspect de la compréhension dans un contexte dialogique.

CHAPITRE 5

Vers un système de dialogue dans le domaine des prescriptions médicamenteuses

Dans ce chapitre, nous détaillons la démarche mise en oeuvre pour la saisie des prescriptions médicamenteuses par le dialogue oral. La section 5.1 introduit ce travail d'un point de vue global avec un exemple qui démontre de façon succincte les étapes à suivre pour la validation d'une prescription liée à un plan de soins. Ensuite, en section 5.2, nous présentons notre analyse de l'interaction conversationnelle qui nous a permis de déterminer les intentions, les stratégies dialogiques et les choix méthodologiques. La section 5.2.2 présentera les actions externes que le système aura besoin de produire pour pouvoir interagir avec un LAP. En effet, un système de dialogue orienté tâche tel que le nôtre nécessite des actions externes tels que la consultation d'une base de données médicamenteuses (BDM) ou d'un plan de soins liés aux dossiers des patients.

Un système de dialogue acquis par apprentissage supervisé nécessite des données d'apprentissage dans un contexte dialogique afin d'apprendre une politique pour répondre aux besoins des utilisateurs. Cependant, dans notre cas, il n'existe pas de dialogue de terrain, il est donc nécessaire de produire des dialogues hypothétiques pour que le système apprenne les actions adéquates à prendre. De plus, dans le domaine de la prescription, le nombre d'attributs est important. Le système doit sélectionner les informations nécessaires pour la validation d'une prescription et amener le prescripteur vers une prescription valide et conforme d'un point de vue des réglementations de la HAS. Pour *'bootstrapper'* (c'est-à-dire, s'autoamorcer) le système de dialogue, nous avons généré des scénarios dialogiques produits par notre grammaire de génération pour apprendre un premier modèle dialogique afin d'en acquérir d'autres, plus écologiques, par la suite. La section 5.2.3 présentera ces scénarios notamment ceux suivant un dialogue coopératif. Par ailleurs, un système devrait être capable de corriger ou supprimer certaines informations sur une prescription. Également, lorsque le système est connecté à un plan de soins, le prescripteur devrait pouvoir être capable de changer de médicament ou de dose selon les alertes envoyées par le LAP avant de valider l'ajout d'un médicament sur une ordonnance. Cette gestion du flux dialogique est décrite dans la sous-section 5.2.4. Suivant le flux dialogique, nous avons mis à jour nos modèles de compréhension. Cette phase de génération de données de compréhension et de dialogue est détaillée en section **??**.

La section 5.3 présentera l'architecture et les paramètres que nous avons utilisés pour

l'apprentissage du modèle de dialogue. Pour obtenir des données dialogiques et effectuer une collecte de données de façon itérative, nous avons déployé le système de dialogue sur un terminal mobile. La section 5.4 présente cette interface mobile déployée pour la collecte de données et l'évaluation de la pertinence de l'analyse conversationnelle. Ce chapitre terminera par la section sur la première collecte de données dans un contexte dialogique ainsi que son évaluation présentée en section 5.6.

5.1 Démarche globale des étapes du système de dialogue

Pour permettre la saisie de prescriptions médicamenteuses de manière orale nous présentons un système de dialogue orienté tâche permettant d'acquérir des informations de manière incrémentale afin d'obtenir une prescription valide selon les réglementations en vigueur. Cette approche par dialogue est nécessaire par le fait qu'une prescription contient un nombre d'informations implicites qu'il faut rendre explicites et faire valider par le prescripteur.

(1)	Ofloxacine	200	mg	2	injections	par	jour
	INN	d-dos-val	d-dos-up	dos-val	dos-uf	O	O

FIGURE 5.1 – Exemple de prescription incomplète alignée avec la sémantique

La Figure 5.1 montre un exemple de prescription dans lequel la voie d'administration du médicament n'est pas précisée. Dans cet exemple, la voie d'administration peut être déduite à partir de la connaissance du médicament et du terme 'injection' qui implique une voie *intraveineuse*. Par ailleurs, le système de dialogue devrait demander à l'utilisateur de préciser une durée pour la prescription. En effet, une prescription émise dans son environnement écologique porte une quantité d'informations non-exhaustives sur le médicament. Ainsi, l'approche que nous proposons utilise le dialogue pour vérifier et compléter les informations de prescription en s'appuyant sur un modèle robuste de compréhension et sur des bases de connaissances métiers.

Définir et mettre en oeuvre un système de dialogue oral pour la saisie de prescriptions médicamenteuses nécessite non seulement un travail de recherche sur l'interaction langagière mais également la prise en compte de l'écosystème métier dans lequel ce système doit s'insérer. Suite à des entrevues régulières avec des experts du domaine, nous sommes partis du constat qu'une prescription médicale orale est régulièrement incomplète et que beaucoup d'informations sont présentes implicitement. Afin de gérer l'implicite, notre démarche a consisté à utiliser les bases de connaissances métiers telles que les bases de données médicamenteuses et un LAP certifié par la Haute Autorité de Santé (HAS). Ces systèmes métier permettent de guider un système de dialogue pour gérer l'incomplétude via des requêtes à l'utilisateur. La Figure 5.2 présente un exemple de dialogue qui illustre cette stratégie.

La Figure 5.2 complète l'exemple de la Figure 5.1 en montrant les étapes qui pourraient

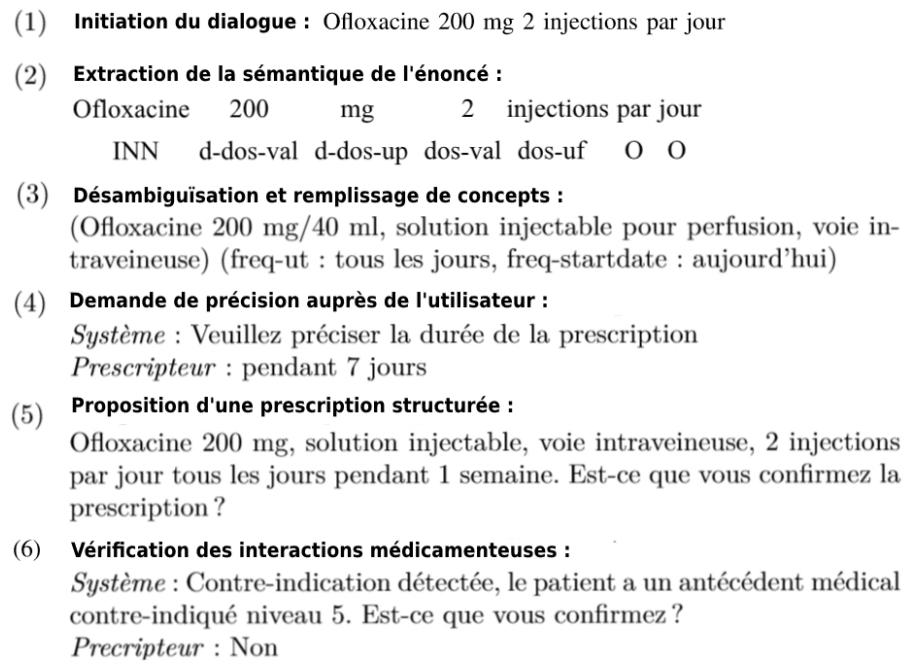


FIGURE 5.2 – Exemple des étapes globales d'une prescription médicamenteuse

constituer une saisie orale. En étape **(1)**, le dialogue est initié par l'utilisateur avec un énoncé oral qui doit contenir le nom d'un médicament. Les valeurs portées par les autres attributs sont également extraites. L'étape **(2)** s'attache à associer l'énoncé à un médicament en trouvant son code d'unité commun de dispensation (UCD). Si aucun médicament n'a pu être associé à partir des attributs extraits, le système suggère de relancer le processus de prescription. S'il y a plusieurs médicaments correspondant aux concepts de l'état actuel du dialogue, le système renvoie une liste de médicaments parmi lesquels le prescripteur doit en choisir un. L'étape **(3)** poursuit le dialogue jusqu'à ce qu'il n'y ait plus informations manquantes pour le LAP. En phase **(4)**, la prescription complète doit être vérifiée par le prescripteur. Des modifications (correction, ajout, suppression) doivent être possibles lors de cette étape. Après la vérification, les données de la prescription sont envoyées au LAP **(5)**. Selon le dossier du patient, le LAP donne des informations sur les interactions médicamenteuses, les allergies du patient, etc. jusqu'à ce que le prescripteur valide la prescription ou l'annule.

5.2 Modélisation du système de dialogue

Dans le chapitre 4 sur le NLU, nous avons distingué les intentions comme étant soit des prescriptions médicamenteuses soit des énoncés hors domaine. Du point de vue médical, nous avons comparé la sémantique des différentes catégories de prescriptions et nous avons cerné le domaine sur cette catégorie majeure des prescriptions médicamenteuses tout en reconnaissant qu'il y a une sémantique commune entre certains types de prescriptions présentée dans l'annexe A. En revanche, lors d'une échange conversationnelle, seule les prescriptions complètes ne seraient pas suffisantes pour arriver à la validation d'une prescription. Pour modéliser les intentions du point de vue de l'utilisateur et le système, nous nous sommes concentrés sur la validation des informations sur la prescription.

Au niveau global, nous avons établi trois points de vérification avant qu'une prescription soit considérée comme étant valide avant l'envoi au LAP. La première phase de la vérification se fait par l'association des médicaments à des codes nationaux d'identification après une étape de désambiguïsation. Cette étape sera détaillée dans la partie 5.2.1. La deuxième étape de vérification consiste à recueillir les informations manquantes à partir des tours de dialogues des prescripteurs via le système de dialogue. Comme nous l'avons évoqué dans la partie 5.1, le nombre de concepts à capturer sur une prescription est vaste. Il faut donc établir les informations obligatoires avant d'envoyer à un LAP. Dans cette deuxième étape, le système cherche à compléter les informations suivantes si elles sont manquantes : la dose du médicament destiné au patient, le rythme de la prescription, la fréquence de cette prise et la durée de cette dernière. Dans cette étape, le médicament est déjà associé à un code d'identification (UCD). Cependant, cette première validation n'assure pas la cohérence d'une prescription. L'exemple suivant présente un exemple de prescription incohérente qui peut-être envoyé au LAP :

> - efferalgan© 500 mg, 1 comprimé à prendre à jeun, le midi et soir pendant le repas pendant 14 jours

L'exemple ci-dessus montre un exemple absurde mais qui est valide en termes d'informations sémantiques. La forme galénique du médicament n'est pas explicitement précisée, et en réalité il se vend sous forme de granulés en sachet ou comprimés orodispersibles. Cependant, la dose prescrite pour le patient (1 comprimé) sous entend qu'il s'agit de comprimés et pas de granulés. Il y a une rythme (matin et soir) et une durée totale pour la prescription. Comme la fréquence de la prise n'est pas explicitement précisée, le système pourrait supposer qu'il s'agit d'une prise tous les jours. La même prescription avec la précision de toutes les informations pourrait être réécrite de manière suivante :

> - efferalgan© 500 mg **comprimés**, 1 comprimé à prendre à jeun, **1 comprimé** le midi et **1 comprimé** le soir pendant le repas, **tous les jours**, **à commencer dès maintenant**, pendant 14 jours

Cette complétion d'information est réalisée par le système avant l'envoi de la prescription au LAP mais cela ne vérifie pas les incohérences qui peut avoir lieu entre les rythmes, fréquences ou les durées. En partant du principe que le prescripteur devrait vérifier la prescription avant et après l'envoi au LAP, cette vérification est réalisée par ce dernier.

La Figure 5.3 situe les étapes précédentes dans leurs interactions avec les bases de connaissances et le LAP. Le système de NLU est sollicité pour chaque entrée du prescripteur. Le *Custom Action Serveur* est le module chargé d'exécuter les traitements qui doivent accéder à une ressource externe. Par exemple, la désambiguïsation des noms de médicament nécessite l'accès à la base de données médicamenteuses Thésorimed®. Le LAP (ici Futura Smart Design®) est sollicité uniquement lorsque l'ensemble des informations obligatoires

d'une prescription sont renseignées. L'une des fonctionnalités majeures d'un LAP est celle de la vérification des interactions médicamenteuses ainsi que la compatibilité du médicament avec le profil du patient. Un LAP peut ainsi renvoyer des alertes auprès des prescripteurs en précisant leurs motifs. À l'étape **(6)**, les données de la prescription sont analysées par le LAP qui renvoie en **(7)** soit une validation, soit une erreur, soit des contre-indications. Celles-ci sont formulées au prescripteur à l'étape **(8)** qui peut ensuite, soit recommencer le processus, soit passer outre les contre-indications en les justifiant par une note (étape non représentée ici).

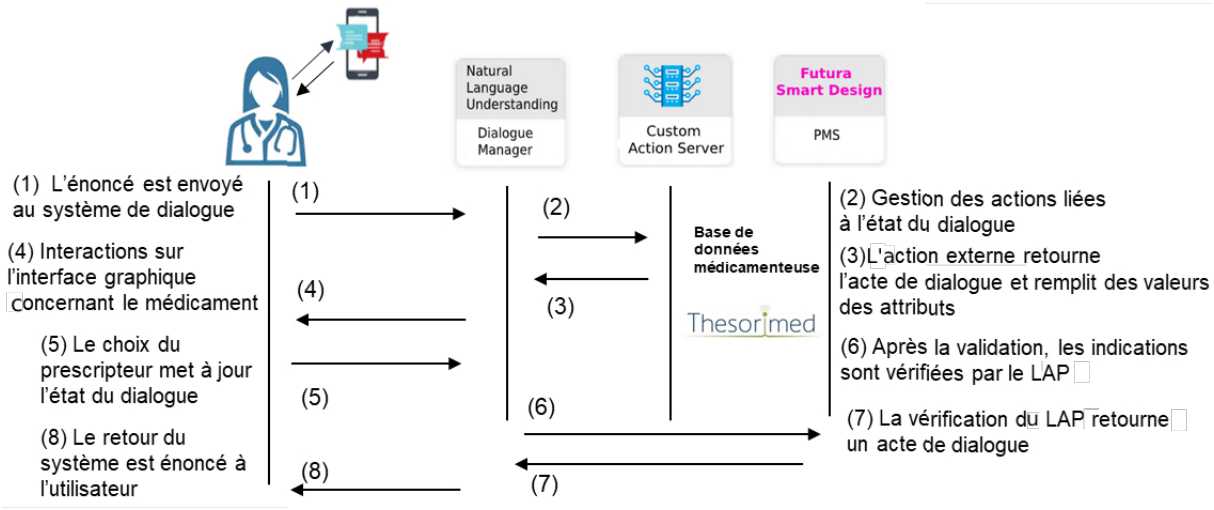


FIGURE 5.3 – Schéma du flux d'information entre les différents composants

5.2.1 Désambiguïsation des noms de médicaments

L'information centrale d'une prescription médicamenteuse est évidemment le médicament. Cependant, identifier de manière unique un médicament est une tâche ardue car le nom ou la molécule n'est qu'un élément permettant de l'identifier. En effet, la liste complète de médicaments utilisés en pharmacie et reconnus par la sécurité sociale décline un médicament selon le principe actif mais également selon la forme galénique (cachet, ampoule...), le dosage, la quantité, etc. Or, dans la pratique, les prescriptions ne sont pas décrites de manière exhaustive et dépendent beaucoup d'un médicament à un autre. Par exemple, pour la prescription d'un médicament où il n'y a qu'un dosage unique comme *Celluvisc*®, la seule information dont nous avons besoin est le nom du médicament. Cependant, pour *Doliprane*® *500 mg comprimés*, même si nous avons plus d'information sur le médicament, il n'y a toujours que deux alternatives : comprimé normal ou comprimé effervescent.

Ce manque d'exhaustivité n'est pas problématique pour un pharmacien qui inférera sans problème l'implicite de la prescription. Mais, dans le cas où la prescription est entrée directement dans un LAP, le prescripteur sera forcé de sélectionner une entrée dans la liste complète de médicaments. Chaque élément de cette liste est associé à un code utilisé dans les bases de données médicamenteuses qui est un identifiant unique de médicament commercial.

Dans notre cas, pour un énoncé oral, nous proposons un processus de désambiguïsa-

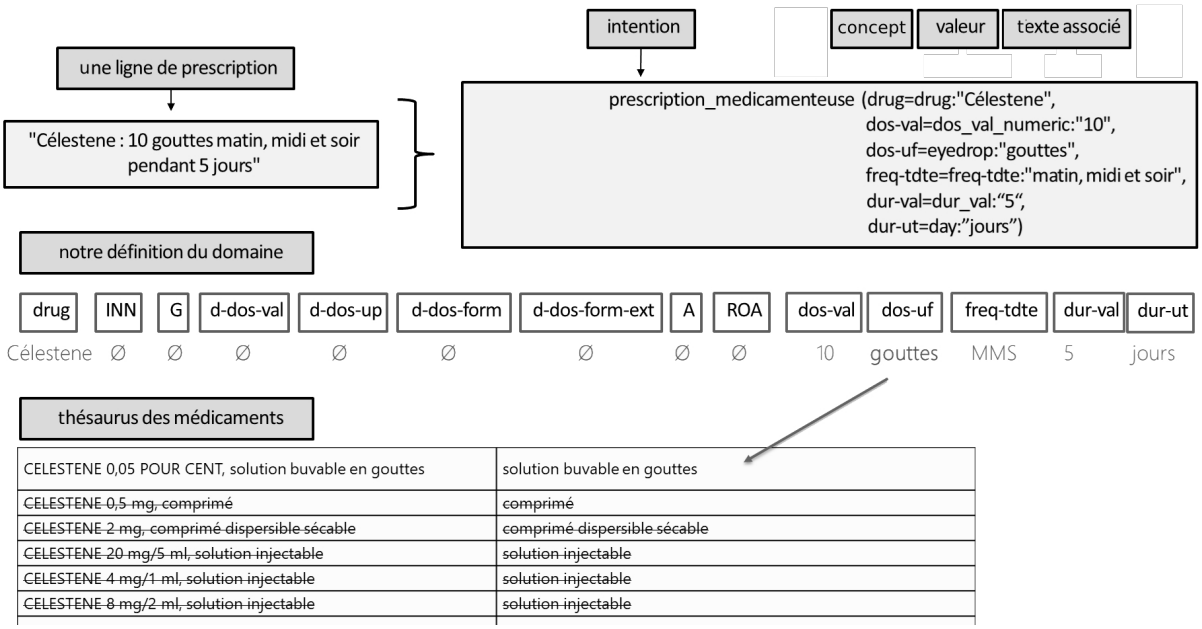


FIGURE 5.4 – Exemple de désambiguïsation d'un médicament

tion utilisant les informations des concepts déjà acquis dans l'énoncé en cours de traitement. Ces concepts sont utilisés pour calculer une similarité avec les entrées de la base Thésorimed® développée par l'Assurance Maladie autorisée par la HAS. L'API développée permet d'associer les concepts à un code UCD (Unité Commune de Dispensation) unique, s'il n'y a qu'un médicament associé, ou à une liste de spécialités avec leurs codes UCD et d'autres attributs sinon. Le processus de désambiguïsation fonctionne à base d'automates (c'est-à-dire des expressions régulières) et de manière itérative en faisant une requête avec le nom du médicament ou de DCI. Si le résultat donne plus d'une réponse alors la requête est étendue avec la dose, puis avec la forme galénique, puis avec la voie d'administration, etc. en fonction d'attributs disponibles. Si aucune réponse n'est trouvée, l'utilisateur est invité à recommencer. Si une seule réponse est trouvée, le système passe à la suite du dialogue. Si plusieurs réponses sont retournées alors l'utilisateur est invité à choisir un élément dans la liste des médicaments possibles.

La Figure 5.4 illustre cette stratégie avec un exemple de désambiguïsation lorsque l'énoncé est "Célestene : 10 gouttes matin midi et soir pendant 5 jours". L'information Célestene®(drug) est associée à l'information sur la forme du médicament (gouttes) pour trouver une seule entrée. Dans cet exemple, l'information précisée est suffisante pour désambiguïser le médicament et continuer la session de prescription. Lorsqu'il y a plusieurs médicaments correspondant aux attributs de la prescription, le prescripteur devra choisir un médicament à partir d'une liste.

5.2.2 Actions du système

Dans un système de dialogue, des connaissances spécifiques à un domaine peuvent être encodées sous forme de modèles d'actions. Ces actions sont souvent associées à l'exploitation de ressources d'une base de connaissances ou une base de données. Dans la partie

initiale du prototypage de notre système, nous avons développé certains motifs d'action nécessaires à la recherche de médicaments. Nous avons réparti ces motifs d'actions selon leurs usages : hospitalier ou ambulatoire. En effet, les modèles d'actions lors d'une prescription hospitalière nécessitent de connaître l'état des stocks, la disponibilité du personnel (p. ex. infirmiers), etc. Les modèles d'actions que nous avons définis sont résumés dans le Tableau 5.1.

Action Système		Description
H		
	action-check-drug	Action qui envoie l'état actuel du dialogue à l'API permettant de trouver le médicament dans la base de données Thésorimed
✓	action-get-roa	Action qui envoie la liste des voies d'administrations paramétrées pour le médicament choisi
✓	action-check-referentiel	Action qui vérifie si le médicament choisi est disponible dans le livret thérapeutique de l'hôpital
✓	action-check-prescription-unit	Action qui récupère les unités de prescriptions paramétrées pour le médicament par la pharmacie de l'hôpital
	action-check-secability	Action qui renvoie l'information sur la sécabilité du médicament
	action-send-pms	Action qui renvoie l'état validé de la prescription au LAP
	action-check-observations	Action qui vérifie si la prescription du médicament nécessite la précision d'une motivation obligatoire par le prescripteur
	action-check-alerts	Action qui récupère les alertes (s'il y en a) envoyées par le LAP

H : Actions réservées à l'usage hospitalier

TABLE 5.1 – Modèles d'actions définis dans le système

Les modèles d'actions résumés sur le Tableau 5.1 montrent les actions externes que nous avons définies pour une intégration complète avec un LAP. Les actions réservées à un usage hospitalier nécessitent un paramétrage manuel effectué par la pharmacie d'un hôpital. Bien évidemment, les unités de prescriptions standards et les voies d'administrations préconisées pour un médicament sont connues par les BDM. En revanche, dans chaque hôpital, les pharmaciens assurent la gestion et la sécurité des soins en les paramétrant selon leurs critères.

Par ailleurs, l'intégration d'un LAP nécessite qu'il y ait des profils des patients préenregistrés sur le logiciel pour que les alertes fonctionnent correctement. Pour cette raison, nous avons séparé la partie de l'association de la sémantique à un médicament des autres modèles d'actions. Dans ce chapitre, nous nous sommes concentrés sur cette première partie des actions sans l'interaction complète avec un LAP.

5.2.3 Génération du flux conversationnel

Dans ce système de dialogue orienté par la tâche, le prescripteur doit initier une prescription par un énoncé oral qui indique son *intention* de démarrer une nouvelle prescription. Cette intention est portée par un ensemble restreint d'*actes de langage* (Searle, 1969). Le prescripteur étant un agent cherchant à faire agir le système de dialogue, il convient également de définir les *actions* que le système doit effectuer pour accomplir la tâche.

Pour définir ces ensembles d'intentions, d'actes et d'actions dans le cas où il n'existe pas de corpus de dialogues de prescriptions médicamenteuses, nous avons créé plusieurs scé-

narios de dialogues en langue naturelle que nous avons soumis à un expert du domaine. Le travail avec l'expert a permis de mettre en évidence le fait qu'une prescription peut être complétée, corrigée, annulée et que l'interface ne doit pas être uniquement orale (par exemple, un choix dans une liste est plus rapide en tactile). Ceci nous a permis de définir des ensembles initiaux d'intentions, d'actes et d'actions qui ont servi de base à la transcription des scénarios dans une représentation informatique.

Scénario A	**Scénario B**
*medical_prescription{"drug","d-dos-val","d-dos-up"} **(1)** action_check_drug * inform{"drug", "d-dos-up", "d-dos-val", "d-dos-form", "inn", "roa", "ucd_code" - slot{"d-dos-form"} - slot{"d-dos-up"} - slot{"roa"} - slot{"d-dos-val"} - slot{"drug"} - slot{"inn"} - slot{"ucd_code} **(2)** prescription_form - slot{"dos-uf"} -slot{"dur-val"} -slot{"dur-ut"} - slot{"dos-val"} - slot{"requested_slot" : null} - action_confirm_prescription * inform{"validate"} - slot{"validate" : "validate"} - action_utter_validation	*medical_prescription{"drug","d-dos-val","d-dos-up"} - **(1)**action_check_drug - slot{"d-dos-form" : "d-dos-form"}- slot{"d-dos-up"} - slot{"d-dos-val"}- slot{"drug"}- slot{"inn"} - slot{"roa"}- slot{"ucd_code"} **(2)** prescription_form - slot{"dos-uf"}- slot{"dur-val"}- slot{"dur-ut"} - slot{"dos-val"}- slot{"requested_slot" : null} - action_confirm_prescription * **(3)**negate {"drug"} OR negate{"d-dos-form"}(...) - action_negate_delete_item **(2)** prescription_form - slot{"dos-uf"}- slot{"dur-val"}- slot{"dur-ut"} - slot{"dos-val"}- slot{"requested_slot" : null} - action_confirm_prescription (...)

TABLE 5.2 – Deux exemples de scénarios de dialogue informatisé
Les * indiquent une information entrée par l'utilisateur.

Le Tableau 5.2 donne deux exemples de scénarios informatisés que nous avons conçus. Les scénarios sont composés de quatre sous-parties. Le scénario **A** représente une situation où le prescripteur initialise le dialogue en utilisant un nom de spécialité d'un médicament et son dosage. Ceci est reconnu comme l'intention de commencer une prescription (`medical_prescription`) les attributs contenus dans l'énoncé sont collectés (ici, `drug`, `d-dos-val` et `d-dos-up`). Puis, l'action `action_check_drug` **(1)** appelle la recherche désambiguïsée du médicament dans une base de données médicamenteuses. Si le serveur d'association de médicaments échoue, le système propose de recommencer la saisie de prescription. Dans le cas contraire, le système donne une liste de médicaments dans laquelle choisir. Une fois le choix effectué par l'utilisateur (`*inform(*drug,...)`), l'action `prescription_form` **(2)** consiste à demander les informations manquantes à la validation d'une prescription. Ces informations sont la dose, le rythme, la fréquence et la durée de la prescription en cours. Après collecte de l'ensemble de ces informations obligatoires, l'action `action_confirm_prescription` est enclenchée pour que la prescription soit visualisée sur l'interface mobile et que l'utilisateur valide la prescription acquise. Le scénario **B** démarre de la même manière mais ne nécessite pas une demande explicite du médicament grâce à la désambiguïsation. En revanche, une fois que la prescription est soumise à la validation, l'utilisateur demande de supprimer une information ce qui entraîne l'action `negate` **(3)**. Le système demande ensuite à nouveau cette information obligatoire à la prescription.

Le gestionnaire de dialogue est entraîné à partir de scénarios comme ceux présentés sur le Tableau 5.3. À droite, se trouvent les actes de dialogues, les attributs et les intentions tels que vu par le système. À gauche, se trouvent la représentation en langue naturelle de ces mêmes informations avec l'instance énonciatrice (U : Utilisateur, S : Système). C'est-à-dire

que les scénarios sont des visions abstraites (*scripts*) qui résument un ensemble d'exemples de dialogues particuliers.

Pour générer ces scénarios, nous avons identifié, avec un expert du domaine, 7 cas initiaux qui couvrent des scénarios de base coopératifs comme ceux sur le Tableau 5.2. Ces 7 scénarios ont ensuite été instanciés avec différentes valeurs d'intention et d'attributs. Le nombre d'attributs et le domaine de valeur de ceux-ci étant très vaste, une génération exhaustive n'est pas possible. C'est pourquoi, afin de couvrir un grand nombre de cas dialogiques, nous avons initialisé les scénarios avec des cas de prescriptions du corpus de compréhension. Ainsi, il était simple d'omettre certains attributs pour provoquer des actions de type `ask` ou `check_drug`. En plus de cette génération automatique guidée par les prescriptions, nous avons inclus quelques scénarios permettant la gestion d'erreurs afin de rendre possible la correction des erreurs de RAP ou de NLU. Avec les scénarios de gestion d'énoncés hors domaine, nous obtenons un ensemble de 14,255 exemples.

U	Efferalgan 500 mg	inform(task=pres-drug,drug="Efferalgan", d-dos-val="500",d-dos-up=mg
S	Veuillez choisir le médicament correspondant à votre recherche : * EFFERALGAN 500 MG, granulés en sachet * EFFERALGAN 500 MG, comprimé orodispersible	*action_check_drug() reqalts([drug="efferalgan",...])*
U	*Choix sur application smartphone*	select(drug="efferalgan", ucd_code="123456789"
S	Veuillez précisez la posologie de la prescription	ask(slot=dos-val || slot=dos-uf)
U	1 comprimé matin midi et soir	inform(dos-val="1",dos-uf="comprimé", rhythm-tdte="matin midi et soir")
S	Veuillez précisez la durée de la prescription	ask(slot=dur-val || slot=dur-ut)
S	Efferalgan 500 mg, comprimé orodispersible, 1 comprimé matin midi et soir pendant 1 semaine, à commencer dès maintenant. Vous confirmez?	action_confirm_prescription()
U	Oui	affirm()
S	Prescription ajoutée au traitement	affirm()

(U=Utilisateur, S= Système)

TABLE 5.3 – Exemple d'un dialogue coopératif

5.2.4 Gestion du flux conversationnel

Les scénarios que nous avons présentés dans 5.2.3 sont exempts d'erreurs. Or, dans un système de dialogue, les scénarios dialogiques doivent couvrir des situations où il y a des erreurs de RAP, de compréhension ou d'erreur de la gestion d'état de dialogue. L'une des approches courantes dans le dialogue est d'utiliser les seuils de confiance des prédictions

pour modéliser cette interaction sous forme de mécanisme de repli (*fallback*) permettant de clarifier ou reformuler des énoncés dont le seuil est trop faible.

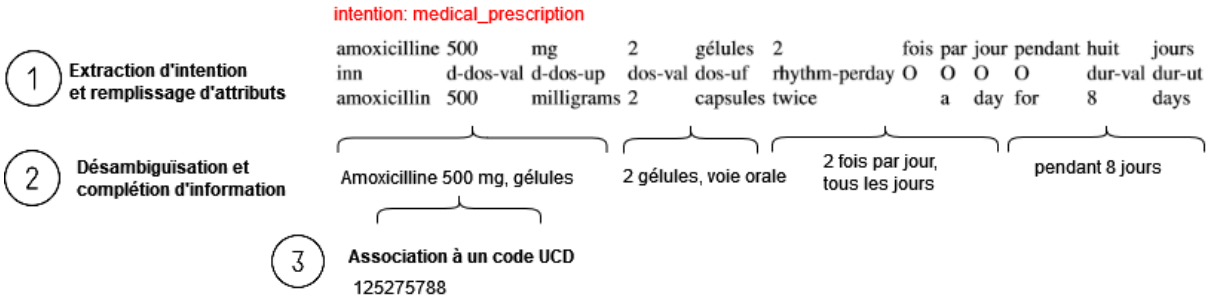


FIGURE 5.5 – Extraction de la sémantique, remplissage d'attributs et association à un code UCD d'une prescription médicamenteuse.

Pour permettre cette gestion, nous avons employé plusieurs stratégies. La première consiste à s'assurer du nom du méedicament. Dans la figure 5.5 nous observons qu'une fois que le prescripteur a initié sa prescription à l'oral, le système essaie d'associer la sémantique à un médicament ou propose une liste de médicaments lorsqu'il y a plusieurs candidats. Si le système n'arrive pas à déterminer un médicament, celui-ci propose de recommencer la session de prescription (*request_restart*). Lorsque le système identifie le médicament et les concepts nécessaires (dose, durée, fréquence, etc.) le système affiche la prescription pour soumettre à la validation du prescripteur. Avant la validation, les attributs identifiés ne sont pas visualisés. La Figure 5.6 montre un exemple de dialogue accompagné de sa visualisation structurée pour validation.

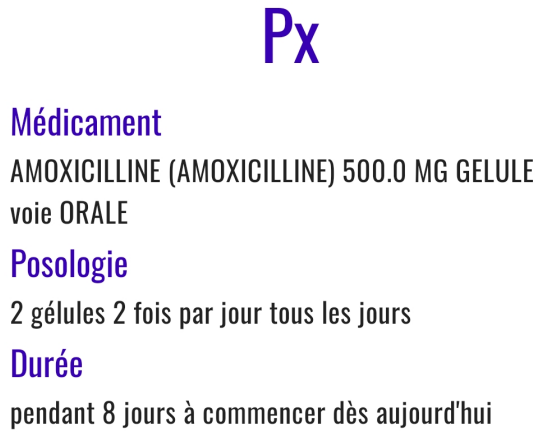


FIGURE 5.6 – Visualisation d'une prescription orale avant la validation sur le terminal mobile

La capture d'écran de l'interface mobile présentée sur la Figure 5.6 montre l'étape avant la validation. Nous avons prévu que la gestion des erreurs (corrections, reformulations) soit faite à cette étape-là. Étant donné que le nombre d'attributs sur une prescription est large, nous avons préféré une visualisation à la place d'une répétition explicite de l'ordonnance qui serait chronophage. Pour la gestion d'erreurs, nous avons ajouté deux intentions : *request_restart* permettant de recommencer le dialogue et *negate* permettant de supprimer une valeur d'un attribut *negate(x)* ou permettant de remplacer sa valeur *replace(x,y)*. Cette stratégie est schématisée sur le Tableau 5.4.

Tour	Intention	negate(x)	replace(x,y)
U	medical_ prescription	Efferalgan 500 mg comprimés 2 comprimés matin et soir tout les 2 jours pendant 14 jours	Efferalgan 500 mg comprimés 2 comprimés matin et soir pendant 14 jours
S	confirm	Est-ce que vous confirmez l'ajout de cette prescription?	Est-ce que vous confirmez l'ajout de cette prescription?
U	**negate/replace**	**Supprimer tout les 2 jours**	**Je n'ai pas dit matin et soir j'ai dit midi et soir**
S	action_negate	(système met à jour les concepts)	(système met à jour les concepts)
S	confirm	Efferalgan 500 mg comprimés 2 comprimés matin et soir pendant 14 jours. Est-ce que vous confirmez l'ajout de cette prescription?	Efferalgan 500 mg comprimés 2 comprimés midi et soir pendant 14 jours. Est-ce que vous confirmez l'ajout de cette prescription?
U	validate	(validation par le bouton)	(validation par le bouton)

TABLE 5.4 – Exemples de dialogues montrant une suppression et une modification d'informations sur des prescriptions

Les exemples de conversations présentés sur le Tableau 5.4 montrent la stratégie de remplacement ou de suppression d'informations. Pour constituer les exemples de NLU, nous avons utilisé les dialogues générés par la grammaire. Pour ce faire, nous avons ajouté des mots clés dans les exemples artificiels tels que « supprimer … par … ».

La suppression devient plus complexe lorsqu'elle concerne des informations obligatoire (p. ex. la dose de la prescription). Dans ce cas-là, le système devrait retourner à l'étape de récupération des informations obligatoires (remplissage d'attributs) et les afficher de nouveau pour soumettre à la validation du prescripteur. Un autre point important concerne la **surinformation** des attributs. C'est-à-dire lorsque le prescripteur précise la valeur d'un attribut déjà connu par le système. Les échanges que nous avons conduits avec les experts du domaine nous a permis de faire la conclusion suivante : les prescripteurs n'ayant pas de temps à perdre, la surinformation pourrait être vue comme une alternative rapide et simple à la correction d'informations. C'est pourquoi, lorsqu'il y a une surinformation d'un attribut, le système remplace la valeur avec la nouvelle valeur et affiche la nouvelle version et demande sa validation. Il faut cependant noter qu'un remplacement d'information au niveau du médicament peut engendrer une nouvelle recherche du médicament dans la BDM et pourrait demander de recommencer la prescription si le médicament n'est pas trouvé. En effet, modifier "doliprane© 500 mg" par "doliprane© 100 mg" désignerait médicament différent.

Dans un premier temps, nous n'avons pas mis en place un mécanisme de repli lié aux seuils de confiance du système de RAP et de NLU. La raison principale de ce choix est qu'il est difficile d'établir un seuil pour une nouvelle tâche (saisie de prescriptions à l'oral), sans se baser sur des dialogues réels. Cependant, nous avons mis en place un bouton « arrière » qui permet d'aller une étape en arrière dans l'état du dialogue en supprimant ce tour de dialogue de l'historique de la pile du gestionnaire. L'expert a suggéré que cela pouvait être pratique lorsque le prescripteur souhaite d'aller une étape en arrière dans sa prescription.

La fluidité du dialogue peut être perturbée par la complexité de certaines prescriptions. Un cas courant de prescription complexe apparaît lorsque le prescripteur formule des prescriptions avec des doses progressives ou dégressives comme dans l'exemple suivant :

- amitriptyline solution buvable à 4%, 10 gouttes le soir pendant 5 jours puis 15 gouttes pendant 5 jours puis 20 gouttes pendant 2 semaines en une prise le soir.

Dans ce type de cas, nous avons prévu une visualisation adaptée qui indique la dose, le rythme, la fréquence et la durée pour chacune des fréquences. Cependant, la gestion d'erreur s'avère plus difficile vu qu'il nécessiterait une analyse temporelle des fréquences. De la même manière, dans certaines spécialités, il est courant de prescrire plusieurs médicaments selon les *degrés d'intentions*. La prescription suivante illustre ce principe des différents degrés d'intentions avec un exemple tiré de Lariven (2008).

Première intention :
— NISIS© 80 mg : 1 comprimé le matin.
Deuxième intention :
— BIPRETERAX© 5 mg/1,5 mg : 1 comprimé le matin.
Troisième intention :
— COKENZEN© 16 mg/12,5 : 1 comprimé le matin.
— AMLOR© 5 mg : 1 comprimé le soir

Dans l'exemple ci-dessus, le prescripteur prescrit le médicament NISIS© en première intention. Cependant, si ce traitement en première intention n'est pas efficace, le médicament BIPRETERAX© devrait être donné comme deuxième intention de traitement. Ce type de formulation dépasse la notion de ligne d'ordonnance. La sémantique de notre domaine et la gestion de dialogue ne prévoient pas des cas de prescription décomposés en plusieurs intentions. En revanche, il pourrait être possible de gérer ce type de cas en ajoutant les médicaments un par un tout en précisant ces degrés d'intentions sous forme de commentaire libre.

5.3 Apprentissage de la politique de dialogue

La politique de dialogue consiste à choisir la meilleure action à entreprendre en fonction de l'énoncé, de l'historique et de l'état courant du dialogue. Dans ce travail, cette politique est basée sur une architecture *transformeur* (Vaswani et coll., 2018) et utilise le mécanisme de *self-attention* (Vlasov et coll., 2019). Plutôt que de considérer l'ensemble de l'historique et les états du dialogue, le mécanisme de *self-attention* permet de concentrer les poids uniquement sur les informations pertinentes par rapport à l'état actuel pour prendre une décision. Cela permet au système de dialogue de mieux gérer les énoncés hors domaine et les scénarios non coopératifs. La figure 5.7 présente cette approche qui fonctionne en cycles. Les attributs (*slots*) actuels, l'énoncé courant et les actions précédentes alimentent un modèle *transformeur* qui génère les plongements d'actions possibles. Il est à noter que dans ce système de *self-attention*, l'attention est portée sur les séquences des tours de dialogues et non sur les termes lexicaux. Pendant la phase d'inférence, le score de similarité de l'état actuel du dialogue est calculé et comparé avec toutes les actions du système (Vlasov et coll., 2019).

À la différence d'une approche classique où le réseau apprend un classificateur pour choisir l'action correspondant à l'état du dialogue, cette approche se focalise sur un classement d'actions basé sur le calcul de similarité suivant $L_{dialogue} = -\langle S^+ - \log(e^{s^+} + \sum_{\Omega^-} e^{S^-})\rangle$.

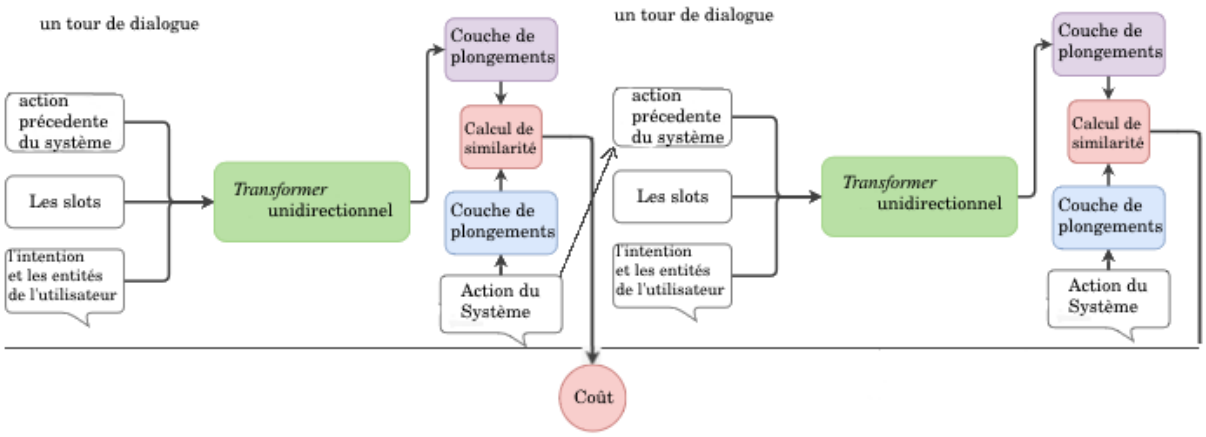


FIGURE 5.7 – Architecture du modèle de gestionnaire de dialogue déroulé sur deux cycles (adapté de Vlasov et coll. (2019))

La sortie du transformeur et l'action de référence y^+_{action} sont plongés dans un espace vectoriel $h_{dialogue}$ et $h^+_{action} \in \mathscr{R}^{20}$ pour obtenir la similarité positive $S^+ = h^T_{dialogue} h^+_{action}$ par produit vectoriel. La similarité avec des actions négatives est obtenue de la même manière $S^- = h^T_{dialogue} h^-_{action}$ avec y^-_{action} une action négative (avec $y^-_{action} \neq y^+_{action}$). La somme est calculée sur l'ensemble des échantillons négatifs Ω^-. Le coût final $L_{dialogue}$ est obtenu par la moyenne de l'ensemble des coûts des cycles. Durant l'apprentissage, $L_{dialogue}$ permet donc de maximiser la similarité positive et minimiser les similarités négatives.

L'entrée du modèle consiste en l'encodage des entrées d'utilisateur, des actions du système ainsi que des attributs. Chaque action, concept et entrée sont encodés sous forme de sacs de mots (Vlasov et coll., 2019). Par exemple l'action du système de vérification de médicament est encodée comme $'action\_check\_drug' = \{action, check, drug\}$. Dans notre cas, 31 actions sont considérées. Concernant les utilisateurs, en plus des entités définies dans la Section 3.3, nous encodons d'autres concepts internes au système tels que le code UCD, l'état de validation de la prescription, etc. L'ensemble d'actions, de concepts et d'intentions donne 147 paramètres. L'architecture *transformeur* est basée sur la configuration utilisée par Vlasov et coll. (2019). La taille des unités du *transformeur* est de 128 avec 1 couche et 4 têtes d'attention avec une dimension de plongement de taille 20. Concernant la régularisation, nous n'appliquons pas de dropout pour l'attention mais un dropout de 0.1 est appliqué pour l'apprentissage des plongements.

Pour l'apprentissage initial, les 14,255 scénarios dialogiques décrits en section 5.2.3 ont été utilisés. Cependant, comme ceux-ci sont générés à partir 7 plans initiaux uniquement, la variation reste faible et la plupart des scénarios sont des dialogues coopératifs. Nous avons donc limité l'entraînement à 20 époques seulement avec une taille de lot de 32 qui est incrémentée de façon linéaire au fil des époques jusqu'à 64. Étant donné la nature de la tâche, nous encodons un historique limité à 10 tours durant l'entraînement.

Étant donné que nous utilisons des scripts de dialogues décrits dans 5.2.3 pour entraîner la politique, le modèle converge rapidement sur toutes les situations à partir de la 12^e époque et obtient une *accuracy* de 0,997. Nous avons choisi le meilleur modèle obtenu à l'époque 16

où la *loss* d'entraînement est la plus basse sur le dev (20% des scénarios dialogiques). Le fait que l'*accuracy* soit très élevée n'est pas surprenant. En fait, les scripts dialogiques permettant d'entraîner la politique du dialogue sont dérivés des cas de prescription standards avec plus ou moins d'attributs sur le médicament et la posologie.

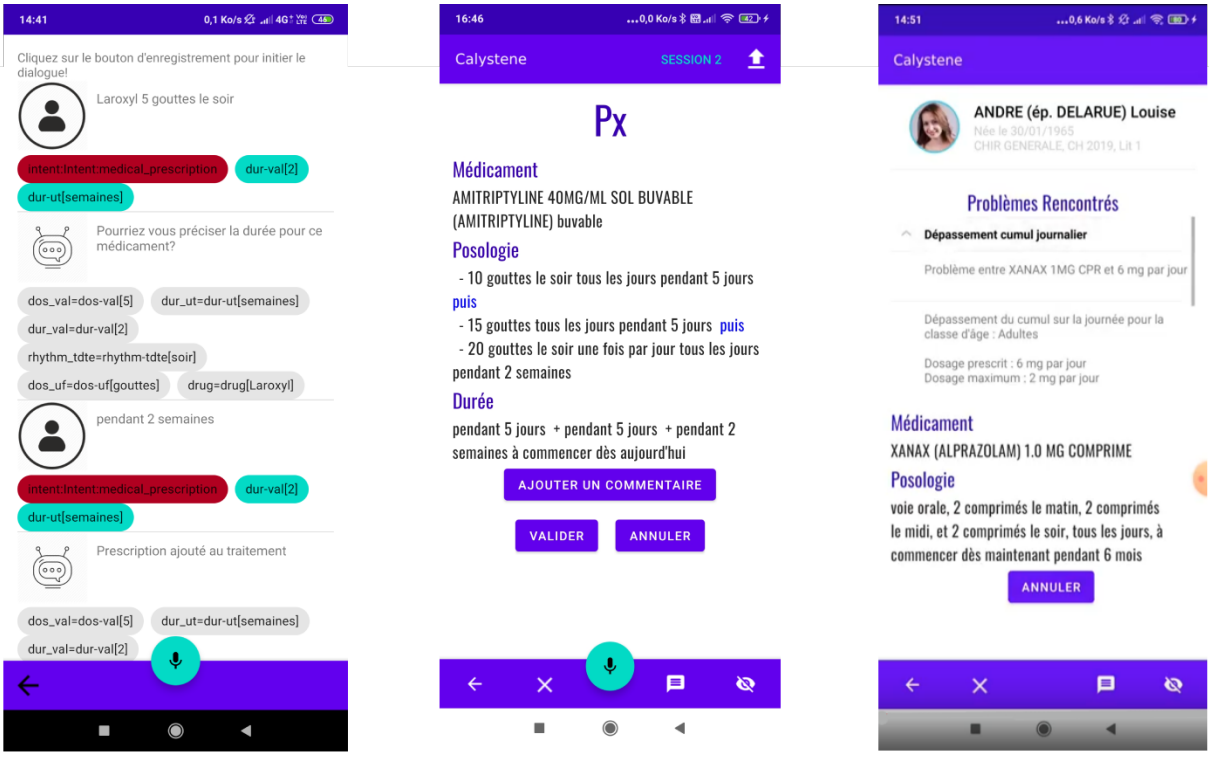
5.4 Interface mobile pour la collecte de prescriptions à l'oral

Pour permettre aux prescripteurs d'effectuer la saisie des prescriptions orales, nous avons développé une interface vocale et mobile sous forme d'application *Android*. Suivant la méthodologie itérative que nous avons établie dans 3.3.4, la mise en place de l'application s'est déroulée en 3 phases :

— Prototype de compréhension du langage : Cette première version était dédiée à la visualisation du remplissage d'attributs qui a évolué au au fur et à mesure en un prototype de dialogue initial.

— Prototype de dialogue : Cette version que nous avons améliorée avec le temps a été développée pour la collecte de données dialogiques en autonomie.

— Prototype intégré au LAP : Cette version connectée au réseau hospitalier d'un LAP a été développée pour effectuer la saisie de prescriptions dans un contexte hospitalier réaliste.



(A) (B) (C)

FIGURE 5.8 – Versions de l'interface mobile pour la saisie de prescriptions orales

La Figure 5.8 montre des captures d'écran de ces 3 versions. **(A)** montre un exemple de prescription enregistré sur le premier prototype développé. Dans cette première version,

tous les énoncés et les inférences internes sont visualisés. Ce prototype nous a permis de valider notre approche avec les experts du domaine et de revoir nos analyses sur l'interaction conversationnelle. **(B)** montre la version que nous avons utilisée pour la collecte de données. L'évaluation qui sera présentée dans 5.6 et la collecte de données réelles présentées dans le chapitre 6 sont basées sur cette version. Dans cette version, les étapes de dialogues suivent le schéma de traitement que nous avons décrit dans 5.1 à l'exception de l'interaction avec un LAP. En effet, l'intégration à un LAP nécessite la mise en place de l'authentification et l'identification d'un prescripteur, le choix des unités fonctionnelles (UF), et d'un patient avant d'effectuer la saisie d'une prescription. **(C)** montre un exemple de saisie de prescription pour une patiente fictive sur la dernière version du système développé après avoir collecté des données. Dans l'exemple présenté dans **(C)**, la saisie du médicament déclenche une alerte dans le LAP indiquant qu'il y a un cumul journalier dépassant les recommandations habituelles.

5.5 Adaptation de notre corpus de NLU dans un contexte dialogique

Comme nous l'avons précisé dans le chapitre 4, nos premiers modèles de compréhension ont été créés à partir des prescriptions complètes. En réalité, pour effectuer une saisie de prescriptions à l'oral, un dialogue est construit au fur et à mesure et doit comporter des exemples pour compléter, corriger ou supprimer certaines informations. Cependant notre approche de génération de prescriptions (cf. section 3.4.2) produit des prescriptions complètes fusionnant des morceaux d'informations liées aux prescriptions. Dans ce contexte, la génération individuelle des morceaux d'informations pertinents constitue une première approche pour générer des exemples. Par exemple, la règle de grammaire ci-dessous indique qu'une posologies est constituée d'une dose à administrer et d'un ensemble de rythmes et de fréquences.

```
PRES[intent=Drug_Prescription,
    RYTHM=?r,FREQ=?f] -> PatientDosage[] DAct_Rythm_Frequency[RYTHM=?r,FREQ=?f]
```

Cette définition de haut niveau de la posologie est suivie des différents types de dosages, de rythmes et de fréquences qui peuvent être remplis avec des valeurs aléatoires. Le tableau 5.5 montre quelques exemples de posologies produits par cette règle de haut niveau rempli avec des valeurs aléatoires.

Même si ces exemples du tableau 5.5 ne sont pas réalistes, ils peuvent permettre au système de compréhension d'inférer l'intention et les attributs des énoncés. Cette génération liée à la posologie permet à l'utilisateur d'informer ('inform') le système sur les éléments d'une prescription. Le tableau 5.6 résume la distribution d'intentions pour le corpus généré par la grammaire.

Le tableau 5.6 montre le nombre d'exemples que nous avons généré concernant les différentes catégories d'informations que nous avons prévus dans le système. Nous avons ajouté

Concepts	Sous-dialogue produit
dos-val, dos-uf, freq-val, rhythm-rec-val,rhythm-rec-ut	95 doses les jeudis toutes les 3 secondes
dos-val, dos-uf, rhythm-tdte, freq-days	548 pilules matin midi et soir les lundis
dos-val, dos-uf, rhythm-rec-val rhythm-rec-ut, freq-val, freq-ut	7 pastilles toutes les 6 heures tous les 7 jours
dos-val, dos-uf, freq-days, rhythm-tdte	zéro virgule vingt neuf litres les jeudis après-midi
dos-val, dos-uf, freq-val, freq-ut, rhythm-rec-val, rhythm-rec-ut	zéro virgule zéro soixante quinze cl toutes les 2 semaines chaque 5 minutes

TABLE 5.5 – Extrait d'exemples sur la posologie générée par la grammaire à traits

Nombre de sous dialogues générés			
inform_drug	700	inform_restart*	95
inform_duration	1196	inform_condition	100
inform_inn	700	request_posology	100
inform_patient_dosage	500	negate(x)	12608
inform_posology	1487	replace(x,y)	12624
inform_rhythm	100	inform_rhythm_freq	400

* : généré pour reconnaître l'intention 'request_restart'

TABLE 5.6 – Récapitulatif des sous-dialogues générés par la grammaire de génération

ces données générées sur le corpus de NLU de référence présentée dans 4.1.3. Nous appellerons cette version du corpus comme **corpus de dialogue** par la suite. Le tableau 5.7 fait une synthèse des intentions d'un point de vue de compréhension après cette génération et le nombre total d'exemples du corpus de dialogue.

Intention	Nombre d'exemples
medical_prescription	8833
request_restart	95
negate	12608
replace	12624
none	1516

TABLE 5.7 – Synthèse des intentions et le nombre d'exemples du corpus de dialogue

Le Tableau 5.7 montre les intentions et le nombre d'exemples pour chacune. La catégorie principale est 'inform_...' liés au remplissage d'attributs des informations sur les prescriptions est résumé dans 'medical_prescription'. Vu que la taille des exemples concernant le domaine des prescriptions a augmenté, nous avons également ajouté d'autres énoncés extraits à partir de ESLO2 (Serpollet et coll., 2007) pour les exemples hors domaine (intention 'none'). Concernant l'intention 'request_restart', elle ne rentre pas dans le flux dialogique que nous avons présenté dans 5.2.4 mais fonctionne comme une intention binaire (*trigger event*) qui une fois reconnue demande explicitement à l'utilisateur de recommencer la session et réinitialise l'état du dialogue si le choix est validé. En termes de nombre d'exemples, ce sont les intentions 'negate(x)' et 'replace(x,y)' pour lesquelles nous avons générées le plus

d'exemples. La raison principale pour générer un nombre important d'exemples de ce type est liée à la variabilité de cas de remplacement d'informations.

5.6 Première évaluation humaine du système de dialogue

Nous avons réalisé une première expérience durant la phase de conception du système afin d'avoir un premier retour. Cette expérience avait un double objectif :

1. collecter un corpus de dialogue oral en français à l'aide du système de dialogue et
2. évaluer le système initial.

Pour cela, nous avons demandé à des experts médicaux et des utilisateurs naïfs d'effectuer des prescriptions médicales en utilisant le prototype. Nous avons préparé deux protocoles pour chaque type de participants. Afin d'éviter que les experts ne se contentent de lire, nous leur avons remis un manuel (Perrot, 2015) présentant des cas cliniques et pour lequel la prescription médicale est présentée sous une forme symbolique ce qui les obligeait à utiliser leurs propres verbalisations. Pour les utilisateurs naïfs, nous avons choisi un manuel destiné aux étudiants en médecine dans lequel les informations relatives aux prescriptions étaient présentées de manière explicite. Elles n'exigeaient donc pas de connaissances médicales pour les énoncer oralement.

Afin de mettre à l'épreuve le système, les prescriptions des manuels ont été classées en fonction de leur complexité. Chaque participant devait effectuer 10 prescriptions en utilisant notre application mobile à l'aide d'un casque et d'un microphone dans une salle sans bruit. Les experts médicaux devaient essayer au moins deux exemples complexes. Au total, 40 dialogues ont été recueillis auprès de 2 experts médicaux et de 2 utilisateurs naïfs. L'évaluation s'est concentrée sur l'acquisition de la prescription et non sur l'interaction avec le LAP.

	Experts médicaux	Utilisateurs naïfs
Taux de réussite de la tâche	45%	16,6%
WER (RAP)	3,40%	17,35%
NLU (micro f-mesure)	75%	43%
Nombre de tours de dialogues sur les vrais positifs	1,56	1,54
Durée moyenne des dialogues sur les vrais positifs	30 secondes	35 secondes
Taux de réussite d'association des médicaments sur les vrais positifs	62%	65%

TABLE 5.8 – Résultats de l'évaluation humaine du système de dialogue

Les résultats sont résumés dans le tableau 5.8. Le taux de réussite global (ratio de prescriptions terminées et validées) est faible tant pour les experts médicaux que pour les utilisateurs naïfs. Pour les utilisateurs naïfs, cela s'explique en partie par des hésitations et une

difficulté à formuler des prescriptions avec un lexique spécialisé. Pour les experts médicaux, les formulations étant plus proches de ceux que nous avons recueillis à partir des livres pédagogiques, le taux était globalement plus élevé. Comme le taux de réussite était faible, nous avons extrait les métriques concernant la durée uniquement lorsque la tâche était réussie (les vrais positifs).

Concernant le taux d'erreurs de mots (WER) de la RAP (ici nous avons utilisé l'API native d'*Android*), celui-ci est bien meilleur pour les experts médicaux, ce qui impacte favorablement la NLU qui présente une F-mesure de 0,75. La situation est moins bonne pour les utilisateurs naïfs avec un WER important et une faible performance de NLU (0,43 de F-mesure).

Un comportement qui est commun aux deux types d'utilisateurs est le petit nombre de tours de dialogues (environ 1,5 tours) Ceci est dû à la désambiguïsation des médicaments car le dialogue s'est fréquemment arrêté suite à l'échec de la recherche du médicament. Sur les vrais positifs, en général, les prescriptions étaient reconnues et transcrites correctement et les utilisateurs n'avaient pas beaucoup d'interaction à faire. Sur les prescriptions qui n'avaient pas été reconnues, vu que la plupart des erreurs étaient liées à l'association des médicaments, le dialogue ne poursuivait pas. Ceci est illustré par le faible taux d'association de médicaments (environ 40 % d'erreurs), et constitue également la cause principale de l'échec de la tâche pour les experts médicaux. Un autre problème était dû à la difficulté de reconnaître la fréquence (par exemple, *toutes les semaines*) et la durée (par exemple, pour *les deux semaines suivantes*). Ces éléments ont également été signalés comme étant difficiles à extraire dans les défis d'i2b2 ainsi que dans Uzuner et coll. (2010a). Cette situation est induite par le manque de données réelles concernant la formulation de la posologie telles que la précision de la durée ou de la dose de la prescription.

Dans tous les cas où le médicament a été correctement associé, la saisie de prescription a pris environ 20 à 30 secondes, ce qui est raisonnable par rapport à une saisie par clavier.

Dans l'ensemble, cette évaluation rapide montre que, bien que certains éléments de la chaîne de traitement des performances satisfaisantes, le traitement global n'est pas assez robuste. L'évaluation de l'architecture modulaire permet d'identifier les améliorations nécessaires telles que la composante NLU (qui doit être apprise avec des exemples contenants des durée et des fréquence de la prescription plus variées), l'identification des médicaments ainsi que la gestion du dialogue. Il convient de noter que la RAP employée est particulièrement performante et ne nécessite donc pas d'amélioration. Enfin, étant donné que le système de dialogue a été appris avec des scénarios uniquement coopératifs, cela explique qu'un nombre conséquent de dialogues sont entrés dans une boucle de repli dès que l'utilisateur a cherché à corriger le système.

L'évaluation a également montré que l'inclusion de non-experts dans le processus met en évidence la différence de comportement et de langage avec l'expert. En effet, les mauvaises performances des utilisateurs naïfs étaient principalement dues à certaines formulations moins techniques et plus familières qui différaient sensiblement de la plupart des exemples d'entraînement. Bien que très limitée en taille, cette évaluation montre l'importance de réaliser des expériences en incluant les utilisateurs cibles.

5.7 Bilan

Dans ce chapitre, nous avons abordé la saisie des prescriptions médicamenteuses dans un contexte dialogique. Notre démarche détaillée dans 3.3 explique comment nous avons défini le domaine et les intentions. Nous avons montré les étapes à suivre dans 5.1 afin d'arriver à la saisie d'une prescription validée par un LAP dans 5.2. Cette étude menée avec un expert du domaine nous a montré qu'il y a beaucoup d'informations implicites dans une prescription qu'il faut désambiguïser. Également, pour assurer la conformité d'une prescription, avant la validation d'un LAP, on doit s'assurer que le médicament en cours de prescription est correctement saisi en l'identifiant par son code UCD.

Pour entraîner une politique de dialogue, nous avons créé des scripts de dialogues (scénarios) pour entraîner la politique afin de revoir la démarche auprès des experts médicaux et des utilisateurs naïfs avant de lancer une collecte à une plus grande échelle. Ces étapes de modélisation et de génération sont présentées dans 5.2.3 et 5.2.4. Pour bénéficier des performances d'une politique de dialogue basée sur l'architecture transformeur (Vlasov et coll., 2019) nous avons entraîné un premier modèle en utilisant ces scripts de dialogue qui couvrent une grande quantité de cas de prescriptions.

L'évaluation humaine que nous avons conduite avec 2 experts du domaine et 2 utilisateurs naïfs présentés dans 5.6, nous a permis de d'obtenir des pistes d'amélioration afin d'établir un protocole pour une collecte de données à plus grande échelle. Le taux de réussite faible de la tâche nous a permis de revoir notre algorithme d'association de médicament, d'adapter l'application mobile selon les retours des experts, de créer des exercices à la fois pour un public naïf et d'experts médicaux, etc. Cette expérience nous a montré que la démarche et la modélisation allaient dans la bonne direction et que les problèmes à résoudre étaient principalement de l'ordre de l'amélioration et de l'enrichissement des exemples d'entraînement.

CHAPITRE 6

Expérimentation du système en autonomie et collecte de données sur terminaux mobiles

Dans ce chapitre, nous présentons une évaluation globale du système à travers une expérience que nous avons conçue pour que les participants puissent tester le système en toute autonomie. Cette méthode nous a permis d'effectuer des expériences malgré les contraintes sanitaires dues à la COVID-19.

La section 6.1 présente le protocole que nous avons mis en place pour la campagne démarrée en janvier 2021, nous en détaillons les modalités et les finalités. Dans cette évaluation, nous impliquons, de nouveau, des participants naïfs et des experts en prescriptions médicamenteuses. Cependant, pour avoir une granularité plus fine, nous demandons aux experts de préciser s'ils sont médecins (utilisateurs cibles du système) ou d'autres types d'experts en prescription (p. ex., pharmaciens).

La section 6.2 présente la phase de préparation des exemples de prescriptions pour ces différents types de publics.

Ensuite, nous présentons les caractéristiques des données collectées en section 6.3.

Les données collectées ont été transcrites et annotées manuellement par des humains. Cette étape de transcription et d'annotation est présentée en section 6.4. Les données collectées nous ont permis de construire un corpus de parole aligné avec les transcriptions et l'annotation sémantique dans un contexte dialogique.

Nous avons utilisé ces données pour évaluer le système. La section 6.5 présente les résultats de cette évaluation.

6.1 Finalités et méthode expérimentale

L'expérimentation que nous avons mise en place a deux finalités principales :

1. évaluer le prototype initial de saisie de prescriptions médicamenteuses et
2. collecter des données pour
   (a) entraîner des modèles de traitement du langage pour améliorer le prototype
   (b) diffuser ces données au sein de la communauté pour permettre la reproductibilité et l'avancée des recherches sur cette tâche.

En effet, étant donné que la plupart des traitements de ce prototype sont acquis par apprentissage automatique, il est essentiel de récolter un corpus suffisamment large et représentatif pour apprendre des modèles de traitement. L'expérience et les données recueillies doivent être représentatives de la pratique des utilisateurs humains, en particulier des experts du domaine. Comme cette pratique n'existe pas, il est nécessaire que les participants puissent au moins l'utiliser aussi naturellement que possible dans leur milieu écologique. L'utilisation avec le minimum de contrainte nous permettrait d'estimer l'utilisabilité du système qui inclut, entre autres, de mesurer le temps nécessaire pour effectuer une prescription. Les données récoltées permettront également d'analyser les stratégies des participants face aux erreurs éventuelles du système. Pour résumer, de nouveaux objectifs annexes ont été définis :

— Valider la pertinence des actes de dialogues prévus dans le système et l'impact de l'IHM sur la saisie de prescriptions à l'oral.

— Mesurer les statistiques liées à la saisie de prescriptions médicamenteuses : le taux de réussite de la tâche, le taux d'erreurs de reconnaissance automatique de la parole, la performance du module de NLU, le taux d'associations correctes d'informations sémantiques aux codes nationaux d'identification de médicaments, etc.

— Collecter un corpus de parole dans une situation réaliste de dialogue via un terminal mobile.

— Distribuer ce corpus de parole à la communauté scientifique et technique en ressource libre (creative commons CC0 1.0 universal - domaine public) après l'avoir analysé et annoté. Ceci permet de favoriser la recherche reproductible et permet le progrès de l'intelligence artificielle dans le domaine biomédical.

— Analyser les stratégies linguistiques utilisées par les participants pour compléter les sous-dialogues liés à la complétion d'informations comme la saisie de la durée totale d'une prescription.

— Comparer le temps de saisie à l'oral des prescriptions médicamenteuses aux ordonnances écrites pour des cas faciles et complexes.

— Tout autre traitement a posteriori sur le corpus acquis pouvant informer la recherche et le développement de systèmes sur les prescriptions médicamenteuses.

Initialement, le protocole envisagé consistait à faire venir les participants au laboratoire dans une salle dédiée à la collecte en présence d'un expérimentateur qui donnerait les modalités, les consignes et les exemples de lecture. La pandémie de Covid-19 a nécessité de changer cette approche pour fonctionner à distance. Nous avons déployé un serveur dédié sous forme d'API permettant de transmettre et récupérer les données à distance. Ceci a demandé un développement et une préparation très conséquents car les participants devaient être en totale autonomie avec leurs propres smartphones pour respecter le protocole sanitaire le plus strict. Notre objectif était d'atteindre environ 50 utilisateurs naïfs et environ 30 experts médicaux, dont une partie de médecins.

Nous avons établi un protocole simple, invitant le participant a suivre les étapes suivantes :

1. S'inscrire sur le formulaire de participation à l'expérience
2. Réceptionner l'apk (fichier d'installation de l'application mobile sur Android) et suivre le document expliquant l'installation et le déroulement de l'expérience
3. Recevoir des exemples de prescriptions (selon le public : 20 exemples de lecture pour les utilisateurs naïfs ; 10 pictogrammes et 10 exemples de lecture pour les experts médicaux.)
4. Participer à l'enquête sur les méta-données (sans information permettant l'identification), accepter les conditions d'utilisation et effectuer l'expérience

Il est important de préciser que l'évolution de l'épidémie de Covid-19 a affecté le déroulement décrit ci-dessus, notamment, pour les médecins hospitaliers. En effet, grâce à une collaboration fructueuse avec le CHU de Grenoble (site Nord) nous avons pu organiser des sessions collectives où au moins un expérimentateur était sur place pour expliquer les consignes et faciliter la mise en place du protocole. La procédure de l'expérience et le guide d'installation envoyé aux participants sont détaillés dans l'annexe B.

En ce qui concerne les données personnelles collectées, après plusieurs discussions avec le délégué à la protection des données (DPO) de l'UGA, nous avons ajouté une enquête au début du traitement de données qui demande les informations suivantes :

— L'autorisation à enregistrer l'audio des interactions vocales en utilisant l'application mobile

— La tranche d'âge (par tranche de 10 ans)

— Le sexe

— Le français est-il la langue maternelle du participant

— Le participant est expert médical, médecin ou aucun de ces cas (naïfs)

Pour permettre la collecte à distance, nous avons intégré ces questions sous forme de formulaire dans l'application de collecte avec une explication des conditions générales.

La Figure 6.1 montre le formulaire sur l'application mobile développée. L'expérience démarre avec ce questionnaire (A). Les métadonnées sont sauvegardées dans une base de données *sqlite* (locale) dans le répertoire de cache de l'application mobile. Pour effectuer la saisie de prescriptions, les participants utilisent le mécanisme du bouton *push* « Appuyer - Parler » qui déclenche l'enregistrement audio, puis de nouveau le bouton *push* pour arrêter l'enregistrement. **(B)** de la Figure 6.1 montre ce principe de fonctionnement. Sur cette capture d'écran, l'enregistrement est démarré et le système enregistre l'énoncé de l'utilisateur. Il est à noter que le prototype initial que nous avions développé ne nécessitait pas le principe d'appuyer-parler et coupait la parole à partir d'un certain seuil de silence que nous avions paramétré. Cependant afin d'obtenir une collecte avec des enregistrements non coupés, nous avons mis en place ce principe d'appuyer-parler. Concernant les mesures d'évaluation que nous avons présentées dans 3.6, l'utilisation de ce principe nous fait perdre l'in-

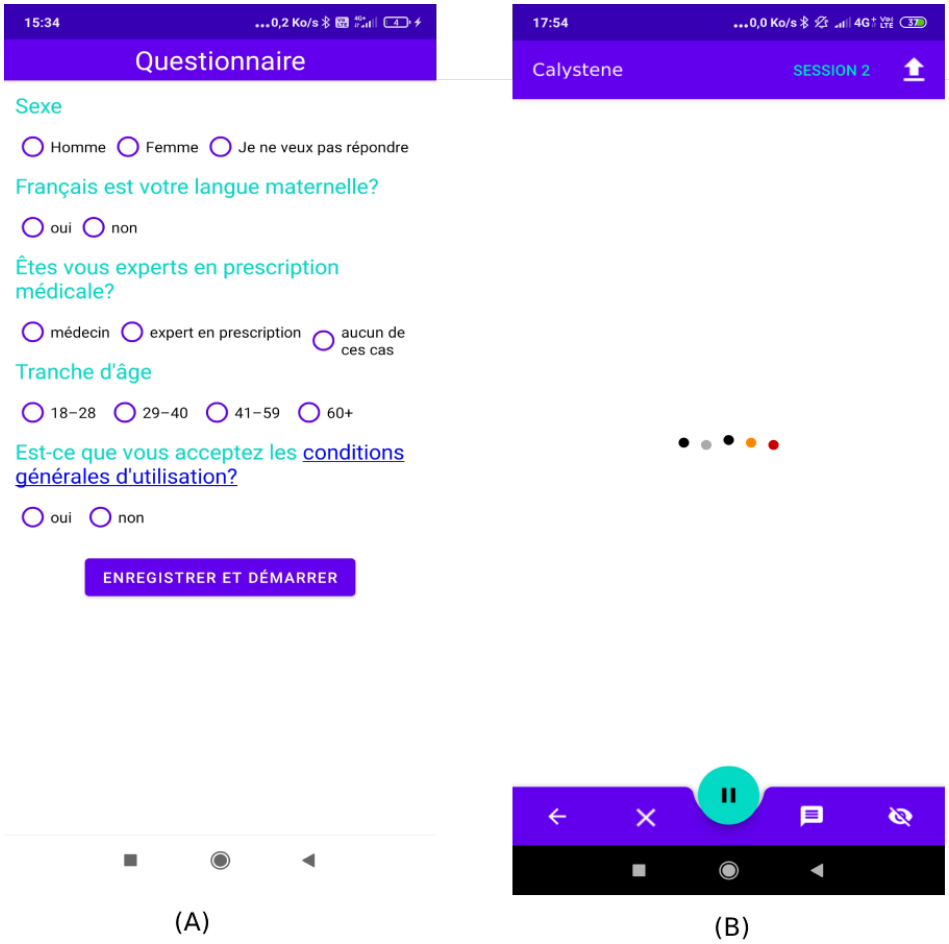


FIGURE 6.1 – Capture d'écran du formulaire des métadonnées (A) et l'écran principal de l'interface (B). Les conditions générales étaient accessibles en cliquant sur le lien.

formation sur les coupures (*barge-in*) et les interruptions qui sont des indicateurs liés à la frustration des utilisateurs, mais permet d'obtenir des enregistrements de meilleure qualité.

Pour la collecte à distance, les enregistrements locaux sont envoyés à notre serveur dédié via une connexion sécurisée (https). L'enregistrement est ensuite analysé par le service de reconnaissance automatique de la parole de Google. Par mesure de sécurité, la transmission des données entre les services de *Google Cloud* est réalisée depuis le serveur et non depuis le portable des usagers. Lorsque le serveur reçoit le résultat de la reconnaissance de la parole (la transcription), celui-ci est envoyé à notre système de dialogue qui extrait l'intention, la sémantique et tente d'associer les *slots* liés aux médicaments au BdM et détermine la suite du dialogue. Le dialogue se poursuit en demandant les informations manquantes ou en répondant aux requêtes de modification de l'utilisateur. Un dialogue est terminé lorsque l'utilisateur valide une prescription complète ou l'annule. La gestion de dialogue comprend les étapes jusqu'à la validation de la prescription par le prescripteur mais ne comprend pas l'intégration d'un LAP.

Tout au long de la collecte, les statistiques des événements et le temps écoulé entre les différentes étapes du dialogue sont enregistrés. Le tableau 6.1 résume les informations que nous enregistrons dans la base de données de la collecte. Toute action effectuée sur l'interface ainsi que l'enregistrement vocal sont enregistrés sous forme d'événement.

Catégorie	Indicateurs
dialogue	identifiant de la session tour de dialogue acteur énoncé/événement timestamp
enregistrements	identifiant de la session tour de dialogue chemin du fichier
métadonnées	tranche d'age identifiant du participant natif_français expert_medical conditions_generales

TABLE 6.1 – Synthèse des informations enregistrées dans la base de données locales de l'application de la collecte

S/U	Info	Tour	Utterance	Timestamp(s)
U	ASR	1	paracétamol 1 g 4 fois par jour pendant 10 jours	40
S	TTS	1	Choissisez le médicament correspondant à votre recherche	46
U	UI	2	listview_item_clicked	54
S	TTS	2	Pourriez vous préciser la posologie pour le patient?	54
U	ASR	3	4 fois par jour pendant 1 mois	94
S	TTS	3	4 fois par jour n'est pas assez précise. Veuillez donner des horaires plus spécifiques.	95
U	ASR	4	8 heures midi 18 heures 22 heures	138
S	ASR	4	Est-ce que vous confirmez l'ajout de cette prescription sur la liste?	139
U	UI	5	/inform{"validate" :"refuse"}	150
S	TTS	5	J'annule la prescription. Cliquez sur le bouton annuler pour confirmer	150

TABLE 6.2 – Informations enregistrées dans la base de données locale d'une session d'enregistrement

Le tableau 6.2 montre un exemple de 'session' de dialogue enregistré de façon horodatée pendant la collecte. Dans cette session de dialogue, on voit que le premier enregistrement commence à la $40^{ème}$ seconde et dure au total 110 secondes, jusqu'au rejet de la prescription. Comme les énoncés, les événements sur l'IHM sont enregistrés avec des mots clés. Par exemple, 'listview_item_clicked' dénote le choix de l'utilisateur d'un médicament à partir d'une liste de médicaments. Comme nous le voyons sur cet exemple, toutes les réponses du système ne sont pas forcément rendues par synthèse vocale (*TTS – Text-to-Speech*), certaines réponses sont données sous forme d'actualisation de l'IHM. Cette information est enregistrée dans la colonne 'Info' de la table. Pour ce corpus, nous considérons qu'un tour de dialogue est terminé lorsqu'une action de l'utilisateur ou du système est effectuée.

À la fin de chaque session de dialogue, l'une des consignes données aux participants était de faire un retour sur la prescription si cette dernière comportait des erreurs (faute d'orthographe, problème d'affichage, erreur de RAP, etc.). Pour cela, nous avons mis en place un bouton 'ajouter un commentaire' qui permet de transcrire directement l'énoncé des utilisateurs sans interprétation. Par ailleurs, le commentaire libre est utilisé également pour ajouter

toute remarque sur la prescription qui n'est pas représentée dans notre espace sémantique (ex. au long cours, pendant la grossesse, etc.)

6.2 Préparation des exemples de prescriptions

Pour l'expérimentation, nous avons préparé un document expliquant comment installer l'application mobile ainsi que les consignes. Ce document est présenté dans l'annexe B. La durée totale de l'expérience que nous avons calculée, de l'installation jusqu'à la transmission de données et la finalisation de l'expérience, est de 30 minutes. Avec cet objectif, en nous appuyant sur l'expérience précédente (cf. section 5.6) et les préparations que nous avons faites en interne, nous avons prévu une vingtaine de prescriptions par participant.

Comme vu précédemment, les deux publics visés (utilisateurs naïfs et les experts médicaux) n'ont pas la même expertise concernant la saisie de prescriptions. Cependant, la participation des utilisateurs naïfs est intéressante notamment pour la variabilité des situations et des stratégies employées pour répondre au système une fois que la saisie est lancée. Pour minimiser les erreurs dues à une mauvaise connaissance du domaine, nous avons favorisé un exercice de lecture de prescription.

Pour les experts médicaux, nous avons préparé des exercices d'oralisation de prescriptions plus naturels et représentatifs de leur pratique qu'une simple lecture. En effet, afin d'éviter que la lecture d'une prescription n'influence les choix grammaticaux du participant expert, nous avons fourni des représentations de prescriptions médicamenteuses sous forme de schémas qui s'approchent d'un plan posologique. Parfois conçus pour les patients, les plans posologiques permettent de représenter les heures et les médicaments pris par les patients avec les conditions et les contraintes associées. Cette forme de représentation utilise des pictogrammes pour représenter visuellement les prescriptions. De cette manière, on peut représenter les éléments d'une prescription en évitant des *priors* linguistiques.

Le pictogramme sur la Figure 6.2 montre un plan de posologie pour la prise du médicament Modopar© . Pour ne pas influencer les prescripteurs avec les jours de la semaine, les jours sont représentés comme (J1,J2,J3,J4,J5,J6 et J7). La forme galénique du médicament est représentée avec une image, qui dans notre exemple représente des gélules mais que certains pourraient choisir d'appeler capsules. Les noms de médicaments, les conditions ou encore les mentions d'administration sont donnés sous forme de texte en haut pour que le prescripteur en prenne connaissance. La posologie est indiquée sous forme d'un calendrier avec des cases qui indiquent les temps des prises et leur continuité dans le temps. La posologie indiquée sur l'exemple ci-dessous dénote la prise progressif d'une gélule du médicament le matin pendant 1 semaine puis une prise matin et soir pendant encore 1 semaine puis une prise matin, midi et soir pendant 2 mois. Un exemple d'oralisation de ce pictogramme peut-être comme suit :

> - Modopar© gélules, une gélule le matin pendant une semaine puis une gélule le matin et le soir pendant une semaine puis une gélule matin, midi et soir pendant deux mois.

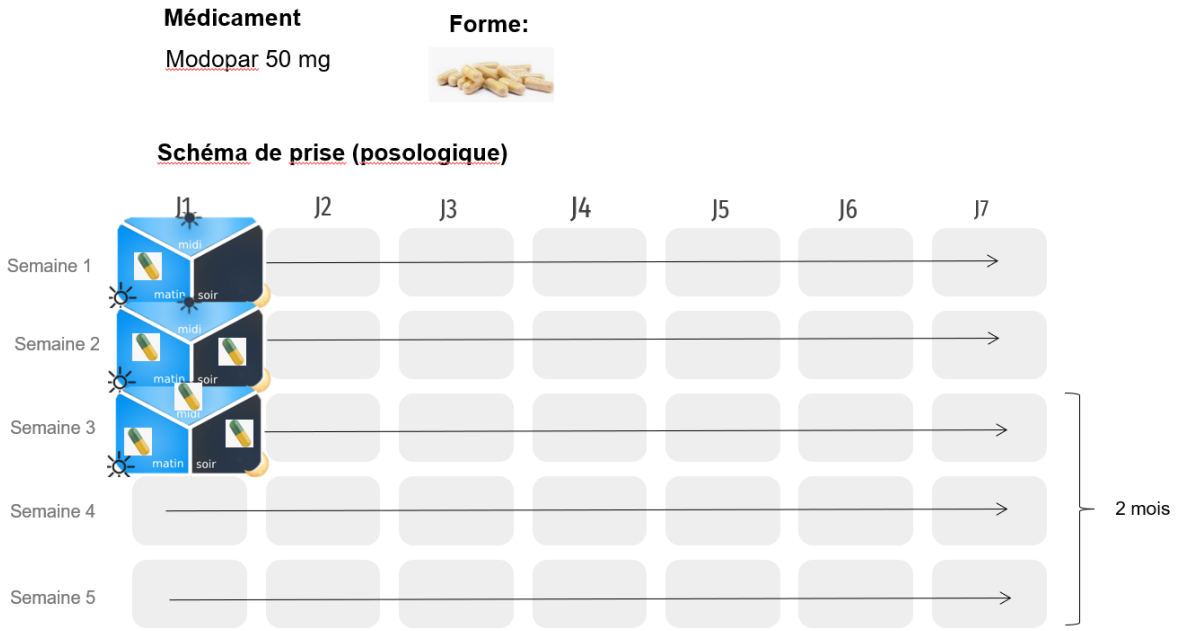


FIGURE 6.2 – Exemple de prescription médicamenteuse sous forme d'un pictogramme représenté par un schéma posologique.

Les pictogrammes présentent un bon compromis entre la lecture d'une prescription déjà rédigée et la définition d'un cas clinique abstrait. Cependant, comparé à l'oralisation d'une prescription écrite, l'exercice prend plus de temps, notamment au début de la passation. Pour ne pas dépasser les 30 minutes définies pour la collecte, nous avons réparti l'exercice entre 10 exemples de pictogrammes puis 10 exemples de lecture. Pour les utilisateurs naïfs, la totalité des prescriptions était des exemples de lecture.

Pour trouver les exemples de prescriptions afin de créer des exemples de lecture et des pictogrammes, nous avons analysé plusieurs livres pédagogiques et thérapeutiques, Schlienger (2013), Delcroix et Gomez (2020), Denis Vital (2018), André (2019) et discuté de ceux-ci avec notre expert. La variabilité était importante, notamment au niveau des fréquences et des complexités de posologie. Afin de faciliter l'apprentissage, les exemples étaient montrés par complexité croissante. Nous avons également filtré ou modifié les exemples lorsque le système de RAP proposait systématiquement des transcriptions erronées pour certains éléments de la prescription (particulièrement les noms de médicaments).

Avec le nombre de participants visé, le matériel généré pour l'expérimentation représente environ 300 exemples de type pictogrammes et 1300 prescriptions médicamenteuses textuelles.

6.3 Caractéristiques des données collectées

La collecte de données démarrée au mois de janvier 2021 a continué jusqu'en octobre 2021 (10 mois d'expérimentation). À la fin de l'expérimentation, le nombre de bases de données que nous avons recueilli est de 55. Ces données contiennent 959 sessions de dialogue

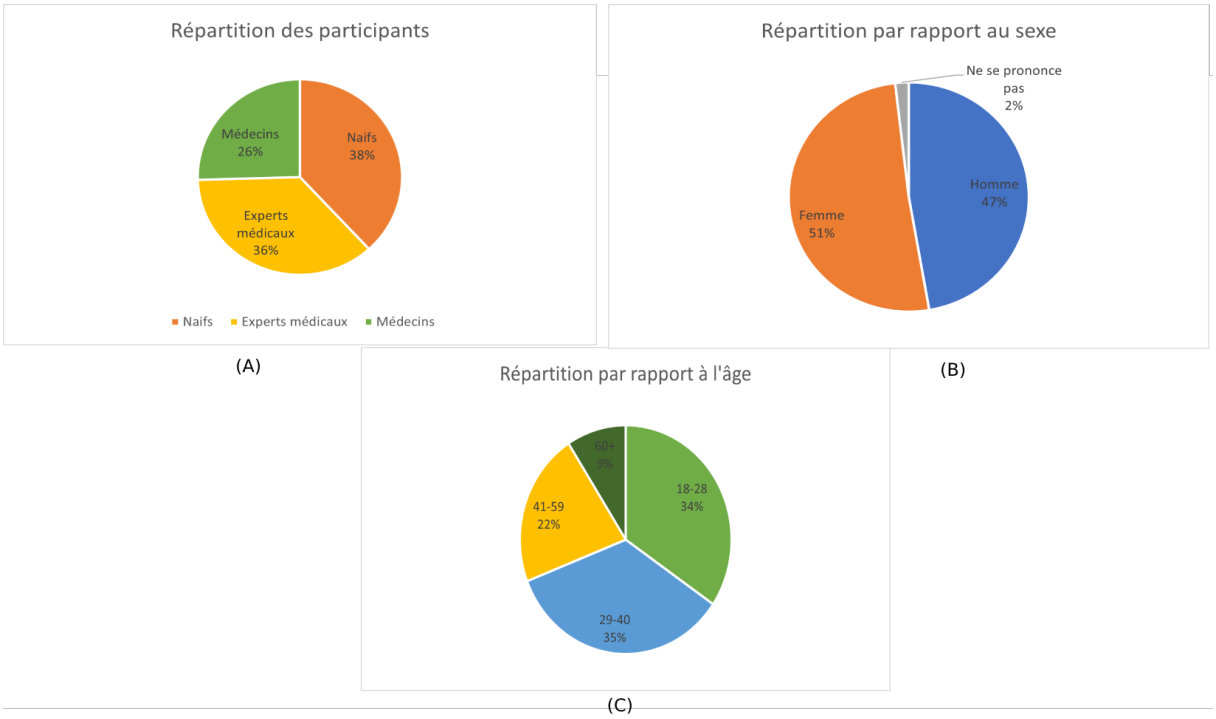


FIGURE 6.3 – Répartition des données par rapport aux méta-données recueillies

comprenant 2067 enregistrements sonores.

Les statistiques sur la collecte présentées sur le Figure 6.3 montrent dans **(A)** la répartition des participants en fonction de leur catégorie : médecin, autre expert médical ou utilisateur naïf. Cette répartition montre que les dialogues recueillis sont équitablement répartis entre les types de participants. Le camembert **(B)** montre la répartition de parité (H/F) du corpus qui montre que les deux catégories sont représentées équitablement. Enfin, **(C)** représente la répartition des données par rapport aux tranches d'âge qui montre que les 3 tranches d'âge principales sont représentées de manière équitable sauf pour les plus de 60 ans qui représente seulement 10%.

	Nb de BD	**Nb de sessions**	**Nb d'enregistrements**	**Temps de parole(m)**
Experts Médicaux	20	279	459	94.83
Médecins	14	246	605	105.21
Utilisateurs naïfs	21	434	1003	62.13
Total	55	959	2067	262.27

TABLE 6.3 – Récapitulatif du nombre de sessions, d'enregistrements et le temps de parole

Le tableau 6.3 présente les statistiques sur le nombre de sessions et les enregistrements sonores de ces différentes catégories de participants. Il montrent que nous avons recueilli 959 sessions de dialogue comportant 2067 enregistrements sonores qui équivaut à 262 minutes de temps de parole soit un total de 4 heures 30 d'enregistrements. Malgré la participation plus faible des médecins par rapport aux autres catégories, ils représentent le temps de parole le plus élevé. Cela peut-être lié aux corrections et ajouts qu'ils ont effectué afin de s'assurer que la prescription soit correcte à leurs yeux. Par ailleurs, le nombre de sessions des participants experts médicaux et médecins est plus restreint car l'oralisation des prescriptions à partir des pictogrammes prend plus de temps que la lecture seule des prescriptions.

Pour rappeler, nous considérons chaque pictogramme ou exemple de lecture comme une session de dialogue.

6.4 Transcription et annotation des données

Les enregistrements que nous avons recueillis pendant la collecte ont été transcrits automatiquement par les services de RAP de Google. Les transcriptions automatiques de Google font parfois des erreurs de transcription et suppriment les disfluences telles que les répétitions, les faux-départs, etc. Ce dernier point est avantageux pour le système de compréhension parce que les énoncés sont en général plus proches des données d'apprentissage. Cependant, afin de refléter la réalité des enregistrements, une transcription plus fine est nécessaire. Par ailleurs, pour l'utilisation du corpus, il est nécessaire d'effectuer l'annotation sémantique. Pour cela, nous avons lancé une campagne d'annotation en recrutant des vacataires chargés de corriger les transcriptions et d'effectuer l'annotation sémantique.

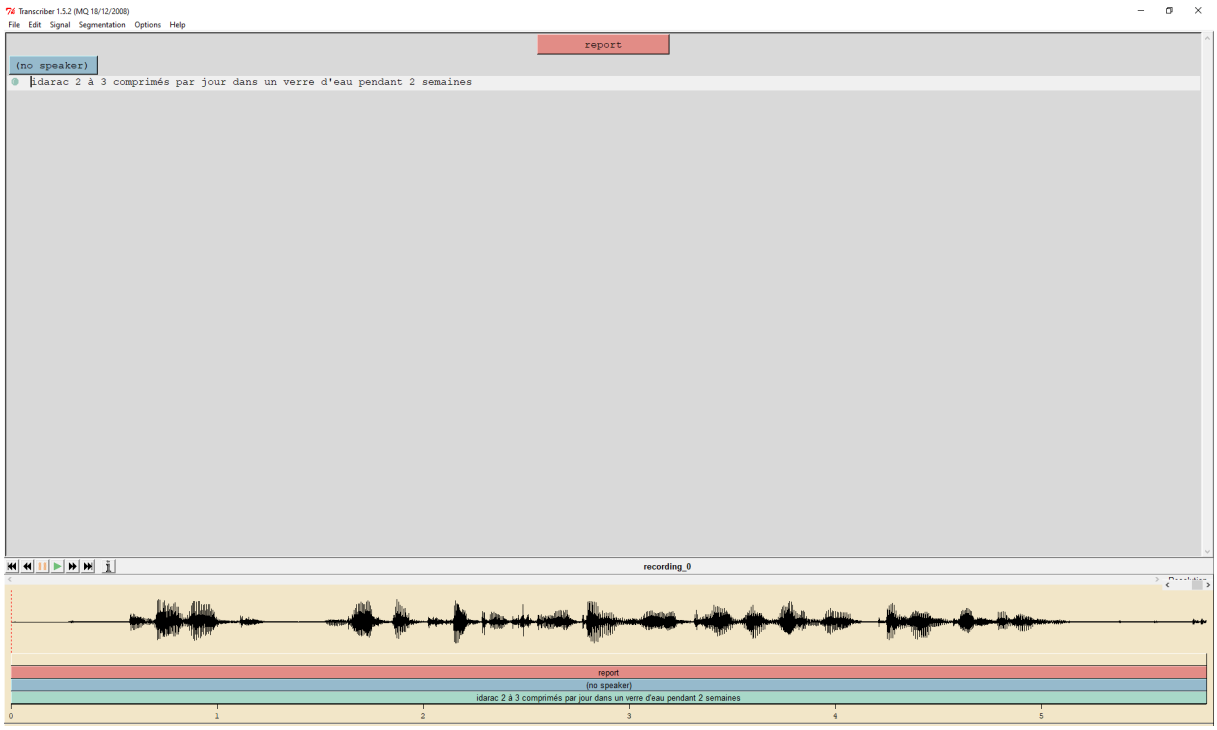


FIGURE 6.4 – Exemple d'une transcription réalisée sur l'outil Transcriber (Barras et coll., 1998)

Pour la transcription, nous avons établi une convention de transcription orthographique. Cette convention est présentée dans l'annexe C. Toutes les transcriptions sont effectuées en minuscules et sans ponctuations. Trois transcripteurs ont transcrit le corpus en utilisant les outils Elan [1] et Transcriber [2] (Barras et coll., 1998). Les enregistrements (sans prendre en compte les sessions dialogiques) ont été extraits dans des lots et répartis selon les transcripteurs.

1. http://icar.cnrs.fr/projets/corinte/confection/elan.htm
2. http://perso.ens-lyon.fr/matthieu.quignard/Transcriber/

Concernant l'annotation sémantique, les trois annotateurs ont utilisé Doccano (Nakayama et coll., 2018) pour vérifier les annotations automatiques produites par notre meilleur système de compréhension (cf. section 4.2). La Figure 6.4 montre l'exemple d'une transcription réalisée sur Transcriber et la Figure 6.5 un exemple d'annotation de NER sur doccano.

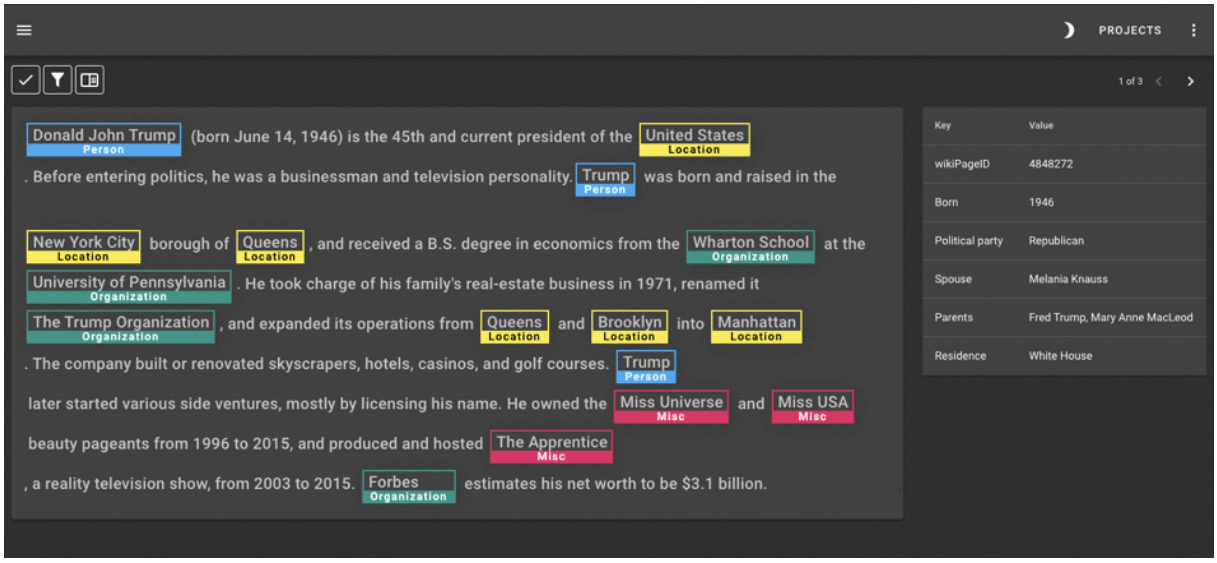


FIGURE 6.5 – Plateforme d'annotation de doccano

Suite à cette phase de transcription et d'annotation, nous avons constitué deux corpus. Le premier est le corpus **PxDialogue** qui contient l'ensemble des dialogues, c'est à dire, non seulement les énoncés utilisateur transcrits et annotés sémantiquement mais aussi toutes les interactions/actions du systèmes. Le deuxième est le corpus **PxNLU** qui est un sous-ensemble de PxDialogue et qui ne contient que les énoncés utilisateur transcrits et annotés sémantiquement (intention et attributs). Le corpus PxNLU permet ainsi d'entraîner un système de NLU ou de SLU en dehors du contexte dialogique alors que PxDialogue peut-être utilisé pour évaluer un système de dialogue.

La figure 6.6 montre un exemple de saisie de prescription (une session) pour illustrer les différences des deux corpus PxDialogue et PxNLU. Dans l'exemple, la partie gauche montre les énoncés de l'utilisateur et la partie à droite montre les réponses du système. Dans cet exemple, après avoir initié une saisie de prescription, l'utilisateur répète la posologie 2 fois (cf. les tours 2 et 3). Il est à noter que nous définissons un tour de dialogue comme le couple consécutif d'un énoncé utilisateur et d'une réponse du système. Par exemple, dans la répartition du corpus PxNLU, l'énoncé '2 fois par jour' ne sera gardé qu'une fois. C'est-à-dire que le corpus PxNLU est composé uniquement de transcriptions d'énoncés uniques. Le tour de dialogue **(5)** de la figure 6.6 est marqué en tant que « commentaire libre ». Expliqué dans l'annexe B, le commentaire libre est utilisé pour ajouter une remarque sur la prescription en cours. Il fait simplement appel à la RAP, mais n'est pas interprété et n'est pas utilisé pour mettre à jour l'état du dialogue.

Les statistiques générales des corpus PxDialogue et PxNLU sont présentées tableaux 6.4 et 6.5.

FIGURE 6.6 – Exemple d'une session de dialogue qui montre la répartition des données en fonction du corpus PxDialogue et PxNLU

Nb de sessions	Nb de tours de dialogue	Nb d'énoncés utilisateur	Temps d'enregistrements
959	3675	2067	262 minutes

TABLE 6.4 – Statistiques générales sur le corpus PxDialogue

Le corpus PxDialogue est constitué de 959 sessions de saisie de prescription. Ces sessions représentent au total 3675 tours de dialogue consécutifs (utilisateur + système) et 2067 énoncés oraux pour une durée totale de 4 heures 22 minutes. Le corpus PxNLU est constitué de 1707 énoncés sur les 2067 total. Le corpus contient 11245 instances d'attribut ce qui représente 6,5 attributs en moyenne par énoncés. Ceci montre que les prescriptions médicamenteuses véhiculent beaucoup d'information sémantique. Le corpus est annoté avec cinq intention dont la répartition donnée dans le tableau 6.6.

Comme dans nos évaluations précédentes, la catégorie principale est celle des prescriptions médicamenteuses qui contient 1545 énoncés. Contrairement au corpus de NLU de référence (cf. section 4.1), ce corpus comprend également des exemples réels sur les intentions 'negate' et 'none'. En plus de ces intentions, il y a une occurrence de demande de recommen-

Nb d'énoncés utilisateur	Nb d'attributs	Nb de tokens	Nb d'intentions
1707	11245	20053	5

TABLE 6.5 – Statistiques générales sur le corpus PxNLU

Intention	Nb d'exemples
medical_prescription	1545
none	111
negate	24
replace	26
request_restart	1
total	1707

TABLE 6.6 – Répartition des intentions du corpus PxNLU

cement de la session exprimée à l'oral. Le Tableau 6.7 présente la répartition des attributs de ce corpus de compréhension.

drug	879	**rhythm-rec-ut**	24	**max-unit-ut**	28
inn	406	**rhythm-tdte**	1291	**min-gap-val**	3
G	1	**rhythm-perday**	239	**min-gap-ut**	3
d-dos-val	815	**rhythm-hour**	104	**cma-event**	251
d-dos-up	807	**freq-val**	16	**fasting**	17
d-dos-form	287	**freq-ut**	99	**rhythm-rec-val**	24
d-dos-form-ext	54	**freq-startday**	10	**re-val**	27
A	43	**freq-days**	18	**re-ut**	25
roa	48	**freq-int-v1**	31	**ns**	0
dos-val	1555	**freq-int-v1-ut**	28	**nr**	0
dos-uf	1475	**freq-int-v2**	21	**qsp-val**	39
dos-cond	117	**freq-int-v2-ut**	12	**qsp-ut**	38
max-unit-val	29	**dur-val**	1182		
max-unit-uf	21	**dur-ut**	1178		

TABLE 6.7 – Répartition des attributs du corpus PxNLU

Comme précédemment, les statistiques sur les annotations présentées sur le Tableau 6.7 montrent une répartition inégale des classes. En dehors des deux attributs sur les remarques pharmaceutiques (ns = non substituable et nr = non remboursable), tous les attributs sont présents dans les corpus. Le nombre de DCI (inn) est également plus représentatif rapport à l'évaluation de NLU présentée en section 4.2.2.

6.5 Évaluation automatique sur le corpus PxDialogue

Nous avons utilisé ce corpus recueilli pour évaluer automatiquement les statistiques sur le dialogue et la compréhension de la parole ainsi que pour les modèles de compréhension que nous avons développés dans le cadre de cette thèse.

Le corpus PxDialogue est composé des transcriptions des sessions de dialogue ainsi que les statistiques que nous enregistrons dans les bases de données locales liée aux sessions.

Ces informations sauvegardées ainsi qu'un exemple de session est présenté dans le Tableau 6.2.

6.5.1 Évaluation de la RAP

Les erreurs de transcriptions ont un impact important sur le module de compréhension et les décisions du système de dialogue. Pour cette raison, nous avons d'abord évalué les performances du système de RAP par rapport aux transcriptions de référence ainsi que la NLU par rapport aux annotations. Le WER de la reconnaissance automatique de la parole est présenté dans le tableau 6.8. Pour faire le lien avec le modèle de compréhension, le Tableau 6.9 résume les performances du système de NLU utilisé pendant la collecte (*Tri-CRF*) qui sera détaillé dans 6.5.3.

	Experts médicaux	**Médecins**	**Utilisateurs naïfs**
WER	21.99%	28.76%	24.42%

TABLE 6.8 – WER des transcriptions automatiques par rapport aux transcriptions de référence

Modèle	**Micro Moyenne**			**Macro Moyenne**			**Moyenne Pondérée**		
	Précision	**Rappel**	**F1**	**Précision**	**Rappel**	**F1**	**Précision**	**Rappel**	**F1**
Tri-CRF	0.83	0.82	0.82	0.64	0.57	0.59	0.83	0.82	0.82

TABLE 6.9 – Moyennes (micro, macro, pondérée) de F1 du modèle

Dans le tableau 6.8, nous présentons le WER des transcriptions automatiques par rapport aux transcriptions de référence. Le *Word Error Rate* (WER) calculé entre les transcriptions automatiques et les transcriptions de référence est aussi élevé pour les utilisateurs naïfs que pour les experts médicaux. Par contre il est beaucoup plus élevé que dans le cas présenté en section 5.6 (4% contre 25%). La raison principale de cette différence est liée à la convention de transcription que nous avons établie. Celle-ci préconisait la transcription des onomatopées, des disfluences, les faux-départs, les répétitions et notamment les troncatures. Il faut aussi noter que dans les transcriptions de référence, les valeurs numériques sont transcrites en toutes lettres à la différence du moteur de RAP que nous avons utilisé. Enfin, l'environnement sonore non-controlé a pu également jouer sur les performances. Le tableau 6.8 ne donne donc qu'un indice sur la performance qui est sous-estimée dues au différences de conventions de transcription indiquées. On peut remarquer qu'il n'y a pas une différence importante entre les utilisateurs naïfs et les experts.

Pour confirmer cette hypothèse sur la baisse des performances liées à la convention de transcriptions établie, nous avons supprimé les disfluences, les normalisations des unités de mesure, les troncatures, et autres automatiquement pour des expressions régulières simples. Le tableau 6.10 montre le WER après avoir effectué ces échanges.

En supprimant les disfluences des transcriptions nous voyons une amélioration d'environ 4% de WER sur les 3 catégories de participants. Ceci confirme qu'une bonne partie du

	Experts médicaux	Médecins	Utilisateurs naïfs
WER	18.67%	24.98%	20.27%

TABLE 6.10 – WER des transcriptions automatiques après l'application des expressions régulières

WER est dû à une convention de transcription différente entre le système de RAP et notre corpus. Le WER reste cependant assez élevé pour ce type de parole (parole proche et qui s'apparente à de la dictée vocale). On peut cependant constater que les erreurs de RAP affectent assez peu la NLU qui reste à un niveau de performance acceptable.

6.5.2 Évaluation du dialogue

Les données dialogiques que nous avons collectées auprès des utilisateurs en utilisant notre système de dialogue dédié au collecte sont constituées de 959 sessions de dialogue comprenant 2067 enregistrements sonores au total. Afin d'estimer la réussite du système de dialogue, nous avons observé plusieurs métriques que nous avons présentées dans la partie 3.6. Les statistiques générales extraites à partir des dialogues produits sont présentées dans le Tableau 6.11.

	Utilisateurs naïfs	Médecins	Experts médicaux
Nb d'enregistrements	1003	605	459
Taux de réussite de la tâche	86%	76%	72%
Temps moyen par session	47,16 (s)	66,15 (s)	35,64 (s)
nb tours de parole moyen par session	3,5	4,28	3,01
nb d'événements moyen par session	7,94	8,44	6,4
Événements système moyen par session	4,01	4,23	3,22
Événements utilisateur moyen par session	3,92	4,21	3,19
Nb d'erreurs fatales	4	4	0
Nb de demandes de répétition	30	32	11
Nb de demandes de restart	33	29	8
Taux d'incompréhension par tour	0.6%	0.10%	0.04%
Taux d'association du médicament	82%	87%	95%

TABLE 6.11 – Résultats des dialogues produits pendant la campagne de collecte de données

Le taux de réussite de la tâche présenté dans le Tableau 6.11 montre que selon tous les

publics, la tâche de prescrire un médicament est réussie en utilisant le système de dialogue à plus de 70%. Le taux de réussite le plus élevé appartient à la catégorie des utilisateurs naïfs. Ceci est sûrement dû à la tâche expérimentale qui consistait uniquement à lire une prescription bien formée. Pour les experts médicaux, la réussite était liée à d'autres facteurs (la compréhension ou non des pictogrammes, de la priorité des informations). Pour les médecins, le taux de réussite de la tâche est légèrement plus élevé que pour les autres experts médicaux.

Un autre indicateur qui est étroitement corrélé avec la satisfaction utilisateur est le temps écoulé pour effectuer une tâche en utilisant un système de dialogue. Le temps moyen de la réussite pour effectuer une prescription à l'oral est d'environ une minute pour les médecins, de 50 secondes pour les utilisateurs naïfs et de 36 secondes pour les autres experts. En lien avec cette durée, on peut remarquer que les médecins et participants naïfs ont plus cherché à corriger les prescriptions que les autres experts. Ce constat est confirmé par le nombre de tour de parole moyen plus élevé pour les médecins et participants naïfs que pour les autres experts.

Nous avons également mesuré le nombre d'événements durant les sessions de dialogue. Comme dans notre cas, l'interaction n'est pas constituée uniquement d'interaction vocale, les événements englobent toute action effectuée sur smartphone (affichage d'une prescription, choix d'un médicament à partir d'une liste, clic sur le bouton arrière, etc.). Nous distinguons les événements du système (ex. affichage) et les événements d'utilisateur (ex. clic sur recommencer la session). On peut constater que le nombre d'événements est corrélé au temps moyen et au nombre de tours de parole par session.

Concernant les métriques permettant d'évaluer la frustration de l'utilisateur, nous avons mesuré le nombre d'erreurs du système, les occurrences de demandes de répétition et *restart* de la part du système. Les erreurs fatales (*crash* système) ont été rencontrées environ dans 4 enregistrements pour les utilisateurs naïfs et les médecins. De même, le ratio d'erreur par rapport aux tours de dialogues est moins de 10% pour toutes les catégories de participants.

Spécifique à notre tâche, l'un des indicateurs pouvant provoquer de la frustration chez les participants est la reconnaissance du médicament. En effet, si le médicament n'est pas reconnu, le système ne donne pas suite au dialogue. Cette situation peut frustrer facilement les utilisateurs qui souhaitent effectuer une saisie de prescription médicamenteuse à l'oral. Le taux d'association du médicament montre que le pourcentage des dialogues où le médicament est correctement trouvé est plus de 80% pour tous les publics. Il est à noter qu'il peut y avoir plusieurs tours de dialogue en cas d'incompréhension pour arriver à ce que le système associe le médicament correctement. Par rapport à l'évaluation humaine présentée dans 5.6, le taux d'association est plus élevé (80% contre 65%). Ceci montre que les corrections apportées à l'algorithme de recherche et d'association de médicaments suite à la première évaluation ont été bénéfiques. Ces corrections techniques comprennent une distinction nette des DCI (ex. Paracétamol), des spécialités (Doliprane©) et des DCI commercialisées (p. ex. Paracétamol MYLAN Pharma©), une meilleure normalisation des noms de médicaments comportant des accents et l'exclusion des médicaments importés dans la re-

cherche lorsque le même médicament se trouve aussi sur le territoire national, etc.

Le taux d'association le plus élevé est celui des autres experts médicaux (95%). Comme ce public est constitué en majorité de pharmaciens, ce taux élevé peut-être expliqué par la clarté de prononciation des noms de médicaments complexes ou les descriptions plus détaillées sur les médicaments qui permettent au système de trouver plus directement le médicament voulu.

Afin d'avoir des mesures plus spécifiques, nous avons analysé les résultats en fonction de l'âge et du sexe des participants. Le Tableau 6.12 montre les résultats sur les dialogues produits en fonction de leur tranche d'âge.

Tranche d'âge :	**18-28**	**29-40**	**41-59**	**60+**
Nb d'enregistrements	674	578	377	283
Taux de réussite de la tâche	74%	79%	81%	87%
Temps moyen	37.02 (s)	42.92 (s)	65.79 (s)	64.49 (s)
Nb moyen de tours de parole	3.21	3.39	3.60	3.82
Nb moyen d'événements	6.87	7.33	8.15	9.06
Taux d'association du médicament	92%	87%	86%	80%

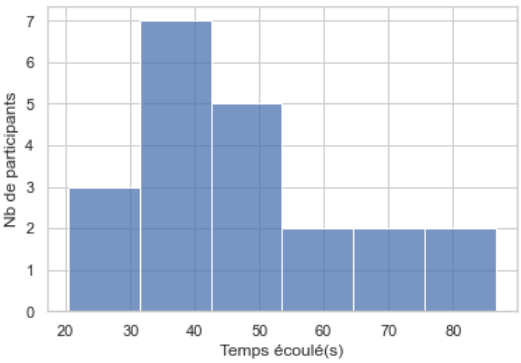
TABLE 6.12 – Résultats avec le système de dialogue en fonction de la tranche d'âge

Le premier constat que nous pouvons faire est que le taux de réussite de la tâche augmente avec l'âge. De même, le temps passé dans une session de dialogue augmente avec l'âge et le taux de réussite. Cela montre que les participants plus âgés cherchent plus à corriger les informations afin d'obtenir une prescription conforme. Pour les autres mesures, le temps moyen, le nombre de tours de parole et le nombre d'événements, ceux-ci augmentent avec l'âge. Nous constatons que les jeunes sont frustrés plus facilement et abandonnent la saisie en cours plus rapidement. Contrairement au taux de réussite de la tâche, le taux d'association du médicament diminue avec l'âge. Cela peut-être lié au fait que 57% des participants de 18-28 ans font partie de la catégorie des autres experts médicaux qui sont les mieux reconnus.

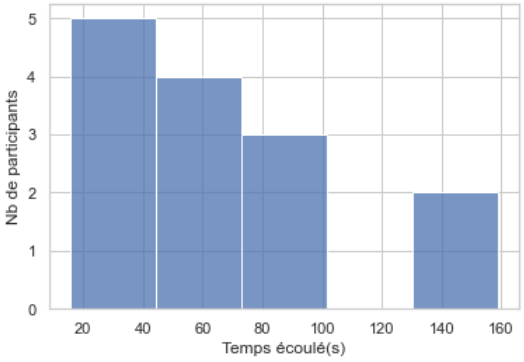
	Homme	**Femme**
Taux de réussite de la tâche	76%	80%
Temps écoulé	51.99 (s)	43.62 (s)
Tours de parole	3.36	3.41
Événements	7.51	7.44
Taux d'association du médicament	89%	87%

TABLE 6.13 – Performances dialogiques en fonction du sexe

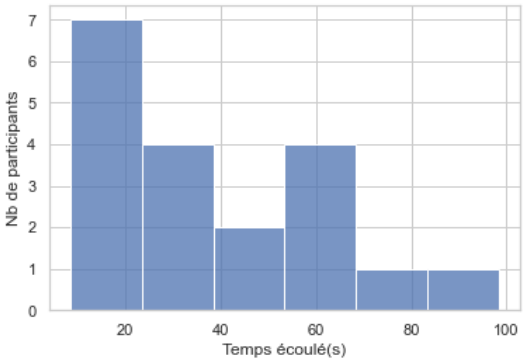
Le tableau 6.13 présente ces mêmes métriques en fonction du sexe. Les résultats montrent qu'il n'y a pas de différence de performances importante entre les 2 sexes. On peut



(a) Utilisateurs naïfs



(b) Médecins

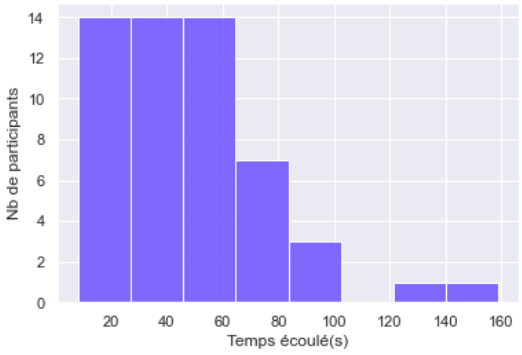


(c) Experts médicaux

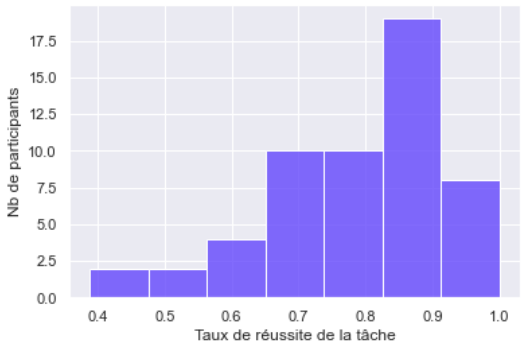
FIGURE 6.7 – Histogrammes du temps écoulé par session pour chaque type de participant

cependant noter que le temps écoulé pour les hommes est légèrement plus élevé que pour les femmes (en moyenne 52 secondes contre 43 secondes).

Les histogrammes de la figure 6.7 montrent les durées de session par catégorie de participant. La durée d'une session est la différence entre les dates du dernier événement et du premier événement de chaque session. Pour les médecins et les utilisateurs naïfs, le temps nécessaire pour effectuer la saisie est un peu moins d'une minute sauf pour quelques participants qui ont mis plus de temps que les autres. Toutefois, pour les autres experts médicaux, nous voyons que la plupart des participants ont effectué la saisie en moins de 40 secondes.



(a) Histogramme du temps écoulé (tous)



(b) Histogramme du taux de réussite (tous)

FIGURE 6.8 – Histogramme du temps écoulé en moyenne en comparaison avec le taux de réussite de la tâche

La figure 6.8(a) montre pour tous les participants, l'histogramme de la durée de réalisa-

tion d'une saisie de prescription (une session). Cet histogramme nous montre que la plupart des participants réalisent la tâche en moins d'une minute sauf pour quelques participants qui ont un délai plus long qui vont jusqu'à 160 secondes, ce qui augmente le temps moyen de saisie sur les métriques.

La figure 6.8(b) montre l'histogramme du taux de réussite de la tâche. Quand nous regardons le taux de réussite de la tâche, nous pouvons remarquer que la grande majorité des participants effectuent la saisie avec 80% de réussite. En revanche, quelques participants ont un taux de réussite beaucoup plus fragile autour de 40%-50% (5 participants sur 55).

6.5.3 Évaluation du module de compréhension automatique du langage

Pour avoir une idée plus claire sur la performance du module de compréhension qui impacte la qualité des dialogues directement, nous avons mesuré les performances des différents systèmes de compréhension présentés en section 3.4. En plus de ces modèles, nous avons également inclus un modèle pré-entrainé pour le français *Flaubert* (Le et coll., 2019) *fine tuned* sur les données d'entraînements du corpus de NLU présenté en section 5.5. Les résultats de détection d'intention ainsi que les scores de précision, rappel et de f-mesure des modèles alignés sont présentés dans le Tableau 6.14.

Modèle	Intention (accuracy)	Micro Moyenne			Macro Moyenne			Moyenne Pondérée		
		Précision	Rappel	F1	Précision	Rappel	F1	Précision	Rappel	F1
tri-crf sur corpus précédent(cf. 4.2.2)	97%	0.90	0.87	0.88	0.68	0.67	0.65	0.90	0.87	0.89
CRF	92%	0.81	0.80	0.80	0.60	0.57	0.56	0.87	0.80	0.83
Tri-CRF	91%	0.83	0.82	0.82	0.64	0.57	0.59	0.83	0.82	0.82
Att-RNN	93%	0.83	0.87	0.85	0.55	0.55	0.53	0.84	0.87	0.85
Flaubert	**94%**	**0.89**	**0.91**	**0.90**	**0.69**	**0.74**	**0.70**	**0.90**	**0.91**	**0.91**

TABLE 6.14 – Tableau récapitulatif des performances des modèles de compréhension sur les attributs-étiquettes du corpus PresNLU

Les résultats sur la prédiction des attributs sont très cohérents avec les premiers résultats de la section 4.2.2. Cela montre que la stratégie de génération de données d'apprentissage a été bénéfique. Le tableau 6.14 montre qu'avec les données réelles, nous avons environ 5% de baisse de f-mesure pour chacun des systèmes.

Les résultats montrent que les performances des systèmes suivent la même tendance que sur les données artificielles. Par contre, c'est le modèle *Flaubert* qui donne les meilleurs résultats quelle que soit la F-mesure considérée : micro, macro et pondérée. Comme nous l'avions précisé dans le chapitre 4, dans notre système, la macro moyenne est très importante puisqu'elle considère tous les *slots* aussi importants quelle que soit leur fréquence. Pour la micro moyenne, on voit une nette contribution du modèle Flaubert.

Concernant les résultats sur la reconnaissance d'intentions, les résultats de tous les modèles sont proches. Cependant, comme le Tableau 6.6 le montre, 90% des intentions collectées concernent les prescriptions médicamenteuses. Pour mieux comprendre, nous avons analysé les scores concernant les différentes intentions dans le Tableau 6.15.

Modèle	***Accuracy* globale**	***Accuracy* sans medical_ prescription**		**medical_ prescription**	**none**	**negate / replace**
CRF	92%	40%	**précision**	**0.96**	0.52	0.50
			rappel	0.98	0.45	0.30
			f-mesure	0.97	0.48	0.37
Tri-CRF	93%	44%	**précision**	0.95	0.56	**0.80**
			rappel	0.98	0.61	0.08
			f-mesure	0.96	0.59	0.15
Att-RNN	93%	48%	**précision**	0.95	0.66	0.62
			rappel	0.98	0.57	0.30
			f-mesure	**0.97**	0.61	0.41
Flaubert	**94%**	**54%**	**précision**	**0.96**	**0.67**	0.79
			rappel	0.98	**0.62**	**0.38**
			f-mesure	**0.97**	**0.64**	**0.51**

TABLE 6.15 – Exactitude (*accuracy*) des modèles pour la reconnaissance d'intentions

Le tableau 6.15 montre l'exactitude globale ainsi que celle calculée en excluant les intentions de prescriptions puis la précision, le rappel et la f-mesure des différentes intentions. L'intention 'request_restart' n'apparaît dans le corpus qu'une seule fois et n'est reconnue par aucun des systèmes. Elle n'est donc pas représentée dans le tableau. L'*accuracy* la plus élevée est obtenue par le modèle *Flaubert*, que les intentions de prescriptions médicamenteuses soit incluses dans le calcul ou non. Quand on regarde les résultats par intention, nous remarquons que *Flaubert* obtient en général les meilleurs scores pour les trois intentions. Le modèle CRF a globalement la précision plus élevée, car le rappel montre que le modèle effectue moins de prédictions erronées. Au contraire, le modèle Tri-CRF a un rappel plus élevé sauf pour l'intention 'negate'. Globalement, l'intention de négation (remplacer ou supprimer une information) est plus difficile pour les systèmes parce que les formulations naturelles de cette intention pourraient être très différentes des données artificielles. Les exemples suivants appartenant à l'intention 'negate' montrent quelques formulations de remplacement ou de suppression.

- ici on dit lévémir et pas levemir et la posologie c'est 10 unités le soir pendant 3 mois
- le magnésium ne doit pas être en gélule mais en ampoule
- problème de compréhension sur la posologie sur le maximum de 2 comprimés

Les formulations en langue naturelle ci-dessus font partie des exemples qui constituent une difficulté pour les systèmes de compréhension que nous avons proposés car contrairement aux exemples d'entraînement synthétiques, ils ne contiennent pas de mot clé explicite indiquant au système un éventuel changement ou suppression d'information (supprimer x, remplacer x par y, etc.). Concernant l'intention 'none', les systèmes ont plus de mal avec les énoncés contenant des informations médicamenteuses mais qui ne sont pas des prescriptions. En effet, le corpus ESLO2 (Serpollet et coll., 2007) permet aux systèmes de recon-

naître les énoncés du français contemporain mais lorsqu'il s'agit d'une conversation liée au domaine médical, vu que les prescriptions médicamenteuses sont majoritaires dans le corpus d'entraînement, ils peuvent être facilement confondus avec les prescriptions. Le Tableau 6.16 montre les scores (précision, rappel et f-mesure) attribut par attribut par le meilleur modèle : *Flaubert*.

Attribut	P	R	F1	Attribut	P	R	F1	Attribut	P	R	F1
drug	0.84	0.88	0.86	**rhythm-rec-ut**	0.71	0.83	0.77	**max-unit-ut**	0.79	0.66	0.72
inn	0.80	0.81	0.80	**rhythm-tdte**	0.95	0.95	0.95	**min-gap-val**	0.21	1.0	0.35
G	1.0	1.0	1.0	**rhythm-perday**	0.91	0.88	0.90	**min-gap-ut**	0.0	0.0	0.0
d-dos-val	0.92	0.93	0.93	**rhythm-hour**	0.74	0.92	0.82	**cma-event**	0.76	0.72	0.74
d-dos-up	0.96	0.94	0.95	**freq-val**	0.74	0.82	0.78	**fasting**	0.18	0.12	0.14
d-dos-form	0.84	0.82	0.83	**freq-ut**	0.87	0.88	0.87	**rhythm-rec-val**	0.56	0.83	0.67
d-dos-form-ext	0.76	0.78	0.77	**freq-startday**	0.57	0.44	0.50	**re-val**	0.63	0.96	0.76
A	0.82	0.98	0.89	**freq-days**	0.31	0.62	0.42	**re-ut**	0.27	0.24	0.26
roa	0.69	0.88	0.77	**freq-int-v1**	0.66	0.81	0.72	**qsp-val**	0.93	0.72	0.81
dos-val	0.90	0.94	0.92	**freq-int-v1-ut**	0.72	0.82	0.77	**qsp-ut**	0.96	0.71	0.82
dos-uf	0.91	0.96	0.93	**freq-int-v2**	0.80	0.76	0.78	**dur-val**	0.96	0.96	0.96
dos-cond	0.55	0.66	0.60	**freq-int-v2-ut**	0.43	0.50	0.46	**dur-ut**	0.97	0.97	0.97
max-unit-val	0.59	0.63	0.61	**max-unit-uf**	0.58	0.64	0.61				

TABLE 6.16 – Précision, rappel et f-mesure de tous les attributs prédits par le système *Flaubert*

Le tableau 6.16 montre que globalement les scores de tous les attributs sont élevés sauf pour quelques attributs qui sont plus difficiles à reconnaître. Si on compare les attributs qui sont moins bien reconnus avec les attributs représentés dans le GuideCorpus (4.1.1, on remarque que les scores plus bas proviennent des attributs qui contiennent en grande majorité des données générées (fasting, freq-int-v1/v2, freq-val, freq-ut, re-val, re-ut, dos-cond, etc.). Malgré le manque de données du domaine, la génération de données permet au système de reconnaître quand même ces attributs, parfois même avec des scores élevés (freq-int-v2, d-dos-form-ext, etc.) cependant il échoue a capturer toute leur variété.

Nous avons voulu tester si le corpus acquis durant l'expérimentation pouvait être utilisé pour le *fine-tuning* de *Flaubert*. Pour cela, nous avons utilisé le modèle pré-entraîné '*flaubert-base-cased*'. L'expérience a été effectuée en utilisant une validation croisée *K-Fold* à cinq tous ($k = 5$). Dans la validation croisée, le jeu de données est découpé en k parties à peu près égales. Dans chaque itération, chacune des k parties est utilisée comme jeu de test et le reste est utilisé pour l'entraînement. Dans chaque cycle, le *fine-tuning* est effectué pendant trois époques. Le tableau 6.17 montre les résultats de cette validation croisée.

K #	Micro Moyenne			Macro Moyenne			Moyenne Pondérée		
	Précision	**Rappel**	**F1**	**Précision**	**Rappel**	**F1**	**Précision**	**Rappel**	**F1**
1	0.92	0.94	**0.93**	**0.80**	0.75	**0.75**	**0.92**	0.94	**0.93**
2	0.92	0.93	0.92	0.77	0.72	0.73	**0.92**	0.93	0.92
3	**0.93**	0.94	**0.93**	0.68	0.65	0.66	**0.92**	0.94	**0.93**
4	0.89	0.90	0.90	0.63	0.53	0.55	0.89	0.90	0.89
5	0.87	**0.95**	0.91	0.71	**0.78**	0.73	0.88	**0.95**	0.91
moyenne	0.90	0.93	0.91	0.71	0.68	0.68	0.90	0.93	0.91
écart-type	0.02	0.01	0.01	0.06	0.09	0.08	0.01	0.01	0.01

TABLE 6.17 – Validation croisée K-Fold (k=5) du modèle *Flaubert* sur le corpus PxNLU

Les résultats sont comparables à ceux donnés dans le tableau 6.14. Cela nous montre

que notre approche sur la compréhension présentée dans 3.4 est valide car l'entraînement avec des données issues des livres ainsi que des données artificielles permettent bien d'obtenir des résultats proches de ceux avec des données réelles. Par ailleurs, les résultats des différents *Folds*(k) varie plus quand on regarde au niveau de la macro moyenne, ce qui est confirmé par l'écart-type qui est plus élevé. Cela montre également que le choix des données impacte la performance macro et donc la couverture des *slots*.

6.6 Bilan/Conclusion

Dans ce chapitre, nous avons détaillé la phase d'expérimentation du système de dialogue que nous avons mis en place. Nous avons présenté les finalités et la méthode expérimentale comprenant la préparation des exercices de lecture et de pictogrammes, l'adaptation de l'application mobile pour prendre en compte une collecte qui peut se dérouler complètement à distance ainsi que la récupération des données des dialogues produits en temps réel.

Dans la deuxième partie du chapitre, nous avons présenté les caractéristiques des données récoltées. En termes de représentation, les données collectées sont bien équilibrées entre les participants experts et naïfs. De même, le nombre d'hommes et de femmes ainsi que la répartition par tranches d'âge sont équilibrés. Au total 959 sessions de dialogues de saisie de prescriptions constituent un corpus de 2067 enregistrements représentant au total 262 minutes d'enregistrement. Nous avons présenté également une convention de transcription et la phase d'annotation manuelle du corpus (transcription et sémantique), ce qui nous a permis de construire un corpus aligné (parole-transcription-sémantique) que nous avons prévu de distribuer à la communauté scientifique et professionnel avec une licence libre de droit.

Les résultats de l'expérience décrite dans ce chapitre montrent que les prescripteurs peuvent prescrire un médicament en moins d'une minute tout en assurant la conformité d'une prescription avec un taux de réussite de plus de 70%. Les résultats sur les modèles de NLU montrent que notre approche pour traiter la contrainte du manque de ressource est valide. Malgré le manque de données réelles dans l'entraînement, les modèles arrivent à obtenir de bonnes performances de prédiction. En outre, la comparaison des différents modèles nous montre que pour la compréhension, le modèle pré-entraîné *Flaubert* est particulièrement efficace. En comparant le *fine-tuning* de ce modèle effectué avec deux corpus différents, le corpus de prescriptions et une partie de PxNLU, nous avons constaté une stagnation des performances. Il semble donc que notre approche de constitution de corpus d'apprentissage permette bien de compenser le manque de données du domaine. Cette approche pourrait donc s'appliquer à d'autres types de prescriptions médicales.

CHAPITRE 7

Conclusion et Perspectives

7.1 Rappel des objectifs et des questions de recherche

La conception d'un système de dialogue dans le domaine médical se trouve au croisement de plusieurs disciplines : le TAL biomédical, le NLU, les systèmes de dialogue, les théories du discours, et bien d'autres encore. Dans cette thèse, nous nous sommes concentrés sur la conception, la mise en place et l'évaluation d'un système de dialogue qui permettrait une saisie de prescriptions médicamenteuses à l'oral. Nous avons abordé la problématique sous forme d'un système de dialogique qui s'intègre à un LAP pour, d'une part, réduire les erreurs de prescription et d'autre part, permettre une plus grande traçabilité des traitements par l'utilisation d'une interface qui serait accessible sur le lieu de soins, potentiellement plus rapide que l'IHM clavier/souris classique.

Dans cette thèse, nous avons cherché une réponse à notre question de recherche principale : **Comment concevoir un système de dialogue couplé avec un LAP dans un domaine n'ayant aucune ressource ?** Pour y répondre, nous avons suivi une approche itérative détaillée dans 3.3.4. En première itération, nous avons analysé des prescriptions médicamenteuses et leurs caractéristiques pour proposer une typologie des informations sémantiques en consultant un expert du domaine biomédical.

L'état de l'art présenté dans le chapitre 2 nous a montré que les systèmes les plus performants sont ceux à base de réseaux de neurones profonds. Cependant, ils nécessitent des données d'apprentissage, en général de grandes quantités, afin d'apprendre des modèles robustes. C'est pour cette raison que nous avions établi la question suivante : **Comment utiliser l'apprentissage automatique pour obtenir des modèles d'inférence (NLU, RAP, politique de dialogue) pour un nouveau domaine qui est a priori dénué de corpus ?**

Afin d'amorcer un système initial, dans la première itération, nous avons extrait des prescriptions médicamenteuses à partir d'un livre pédagogique dans le but de récupérer un premier corpus qui était annoté manuellement. Nous avons ensuite proposé une méthode de génération de données artificielles permettant de créer des données équilibrées qui couvrent tous les attributs définis dans le domaine. Effectivement, la plupart des cas de prescription concernent les informations les plus courantes. En revanche, un système permettant la saisie de prescriptions ne devrait pas se limiter à des cas standards mais devrait également traiter les cas les plus rares, avec une même efficacité. La méthode de génération de données artificielles que nous avons présentée dans 3.4.2 nous a permis d'augmenter

la couverture des attributs qui sont les plus difficiles à trouver sur les ordonnances habituelles. Enfin, avec le premier corpus extrait d'un livre pédagogique et les données synthétiques, nous avons entraîné et évalué différentes approches sur la compréhension automatique dans le domaine des prescriptions médicamenteuses.

Suite à la création des modèles de compréhension, nous avons abordé la saisie d'une ordonnance à l'oral par le biais du dialogue. Notre question de recherche était la suivante : **Comment pourrions-nous modéliser l'échange entre les différents acteurs lors de la saisie des prescriptions médicamenteuses à l'oral alors que cette pratique est inexistante ?** Pour répondre à cette question, nous avons fait une analyse de l'interaction conversationnelle, puis modélisé les intentions et les scénarios dialogiques. En parallèle, en utilisant notre méthode de génération de données artificielles, nous avons généré des données dialogiques permettant de compléter/supprimer/remplacer des informations sur les prescriptions. Le chapitre 5 explique les étapes d'évolution de cette étude revue plusieurs fois avec des experts du domaine ainsi que des utilisateurs naïfs. Nous avons également développé une application mobile permettant d'effectuer la saisie dans un contexte dialogique. Ce premier prototype nous a permis d'effectuer une première évaluation du système de dialogue validant l'approche avant d'entamer une collecte à plus grande échelle.

Nous avons effectué 2 évaluations intermédiaires : une concernant les modèles de compréhension sur les prescriptions complètes dans 4.2.2 et une sur le dialogue avec 4 participants dans 5.6. Cependant, il est difficile d'évaluer la qualité, la pertinence et l'utilisabilité d'un système de dialogue avant qu'il ne soit pratiqué dans une situation réelle. Nous nous sommes donc posés la question : **Comment peut-on mesurer les performances de toutes les étapes du pipeline ainsi que le pipeline dans son ensemble ? Comment mesurer l'utilisabilité de ce système dans un milieu médical ?** En ce sens, la première évaluation humaine avec les utilisateurs naïfs et les experts médicaux nous a permis d'esquisser des pistes d'amélioration pour une évaluation à plus grande échelle.

Notre première intention était d'effectuer une collecte de données auprès des utilisateurs naïfs et des experts médicaux et par la suite d'effectuer une évaluation où les prescripteurs pouvaient effectuer la saisie de prescriptions à l'oral sur le système dans une situation écologique. Cependant, avec la pandémie du Coronavirus, les médecins ont été en première ligne pendant cette crise sanitaire et il n'était pas possible d'effectuer des évaluations sur le lieu de soins. Malgré tout, nous avons adapté le protocole de la collecte de données pour permettre une collecte de données à distance mais finalement se rapprochant au plus d'une situation réelle de prescription, grâce à la mise en place des pictogrammes décrite dans le chapitre 6. Nous avons présenté les résultats de cette évaluation dans la sous-partie 6.5.2, résultats qui ont confirmés l'intérêt de l'approche présentée dans cette thèse.

7.2 Contributions

Dans cette thèse nous avons exploré la saisie de prescriptions orales effectuée sur un terminal mobile via un système de dialogue. Nous avons présenté une étude détaillée des ca-

ractéristiques des prescriptions médicales, notamment médicamenteuses. Notre approche aborde le sujet dans l'optique d'une modélisation de remplissage d'attributs d'un point de vue de TALN biomédical. Cette étude de caractéristiques nous a permis de produire une première contribution au domaine en élaborant une taxonomie exhaustive de la sémantique des prescriptions médicamenteuses exprimée en langage naturel.

Concernant la compréhension automatique de la sémantique des prescriptions, nous avons proposé une méthode pour obtenir des données issues des livres pédagogiques qui donnent des performances raisonnables sur des énoncés réels issus d'une collecte. Nous avons également présenté une méthode de génération de données artificielles qui nous a permis d'augmenter la quantité de données pour équilibrer les classes sous-représentées dans les exemples recueillis. Cette méthode a permis de construire un corpus équilibré en terme de distributions des attributs et suffisamment important pour utiliser des techniques d'apprentissage automatique.

Dans un contexte de faibles ressources, nous avons exploré plusieurs solutions afin d'améliorer la qualité des modèles de compréhension. Nous avons présenté une méthode semi-supervisée et nous avons comparé plusieurs modèles de langue pré-entraînés dans le domaine biomédical sur la tâche d'extraction de médicaments d'I2B2-2009. Les résultats de ces expériences nous montrent que l'utilisation des modèles pré-entraînés est particulièrement performante. La méthode semi-supervisée pourrait être une alternative pour les langues pour lesquelles il n'existe pas de modèle pré-entraîné.

Enfin, nous avons présenté une évaluation du système de dialogue après avoir effectué une période de collecte, de transcription, d'annotation et d'évaluation qui a duré environ un an. Au total, nous avons collecté 4 heures 30 d'enregistrement sonore comprenant plus de 2000 enregistrements. Pour permettre la reproductibilité de la recherche, nous allons mettre à disposition de la communauté ce corpus de NLU et de dialogue.

7.3 Limites et perspectives

Dans cette thèse, nous avons suivi une approche itérative qui nous a permis d'amorcer un système, de valider sa modélisation auprès des experts et d'obtenir des données à la fois pour obtenir des scénarios de dialogues réalistes et des données d'entraînement sur les prescriptions médicamenteuses à l'oral. Ce système a été interfacé avec un LAP ce qui a permis d'obtenir un mécanisme de *feed-back* en temps réel par le biais du dialogue. Cette dernière étape n'est pas présentée dans ce manuscrit, car elle n'a pas été évaluée lors de l'expérimentation. En effet, l'évaluation a porté sur la saisie et la validation d'une prescription par le prescripteur, mais sans lien avec un patient particulier. Or un LAP peut donner des informations plus personnalisées (ex. allergies) ou les contre-indications avec des médicaments en cours enregistrés dans les dossiers des patients. Une perspective directe de ces travaux serait d'évaluer le système dans un contexte écologique complet. Cela permettrait de confirmer les évaluations du système (qui ont été effectuées de manière réaliste, mais qui n'ont pas concerné de vrais cas de prescriptions dans une vraie situation écologique) tout en ajou-

tant la dimension patient. Nous pourrions ainsi vérifier si les interactions supplémentaires, notamment lorsqu'il y a des contre-indications, permettent toujours un temps de saisie raisonnable. Par ailleurs, en lien avec des clients de l'entreprise, des comparaisons entre des saisies via l'interface du logiciel et l'interface dialogique pourraient être menées. Enfin, sur le long terme, l'intégration au LAP pourrait mesurer si l'interface dialogique a un effet positif sur la traçabilité et la réduction des erreurs.

Concernant la compréhension automatique des prescriptions médicamenteuses, les méthodes état de l'art (au début de cette thèse : tri-crf, att-rnn, seq2seq) ont évolué très rapidement vers des modèles pré-entraînés. Avec la disponibilité des modèles de langues, notamment entraînés sur des corpus médicaux, les performances des différentes tâches ont été augmentées significativement. Pour prendre en compte la puissance des modèles transformeurs, nous avons intégré *Flaubert* seulement dans cette dernière version développée après la collecte de données. Il serait intéressant de voir l'impact de ce modèle sur les métriques de dialogues, notamment dans une évaluation écologique.

L'amorçage du système a été réalisé avec succès via l'utilisation des données synthétiques. Si cette grammaire de génération permet d'augmenter la représentation des *slots*, elle reste assez peu représentative de la réalité du terrain. Par ailleurs, la création de règles pour la génération de données est un processus chronophage. Maintenant que nous disposons de données réalistes, il serait intéressant de tester des réseaux antagonistes génératifs (GAN) pour la génération de données de NLU ou de dialogue. Des études récentes Golovneva et Peris (2020) ont pu montrer qu'avec une approche GAN, il est possible de générer des données de NLU même avec des ressources limitées.

Dans cette thèse nous sommes concentrés sur les prescriptions médicamenteuses, mais d'autres prescriptions, détaillées dans l'annexe A peuvent être considérées. Dans l'étude menée en annexe, nous montrons que certains types de prescriptions possèdent une sémantique commune. Il pourrait donc être intéressant d'apprendre des modèles de compréhension sur plusieurs types de prescriptions en même temps sous la forme d'un apprentissage multitâche pour que les connaissances acquises sur un type de prescription puissent se transférer sur un autre type. Ce type de transfert pourrait aussi s'opérer entre différentes langues. Par exemple, il serait intéressant d'explorer des approches bilingues (Anglais/Français) (Rebholz-Schuhmann et coll., 2013) sur des corpus de comptes-rendus médicaux ou des corpus de prescriptions oraux ou textuels. Pour pourrions ainsi exploité le corpus acquis dans cette thèse, mais également le vaste ensemble de données MIMIC en langue anglaise pour transférer des connaissances de l'anglais vers le français.

L'approche modulaire d'un système de dialogue permet de mettre un mécanisme de contrôle sur les processus, apporte une traçabilité des erreurs et permet l'intervention manuelle des ingénieurs. Cependant, elle implique généralement un entraînement séparé des modules ce qui les rend plus sensibles aux erreurs des modules en amont. Certaines approches prônent une fusion de certains modules. Par exemple, Desot et coll. (2019b,a) montrent qu'il est plus avantageux d'utiliser un système de SLU de bout en bout ou les attributs et l'intention sont extraits directement d'un signal de parole, qu'un système pipeline.

Une perspective de recherche intéressante pourrait être d'explorer la question suivante : serait-il possible d'optimiser conjointement tous les modules d'un système de dialogue, tout en gardant des modèles séparés pour chaque module ? Dans l'état de l'art, nous avons présenté quelques approches qui visent une optimisation de modules par un apprentissage conjointe. Par exemple, Zhao et Eskenazi (2016) ont montré qu'il est possible d'entraîner conjointement la gestion de suivi (DST) et la politique du dialogue. Il serait donc intéressant d'explorer ces approches pour à la fois augmenter les performances et conserver la modularité qui est très importante pour la maintenance logicielle.

Bibliographie

ABBOUTE, A., BOUDJERIOU, Y., ENTRINGER, G., AZÉ, J., BRINGAY, S. et PONCELET, P. (2014). Mining twitter for suicide prevention. Dans *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 250–253. Springer.

ABDUL-RAUF, S., KIANI, K., ZAFAR, A., NAWAZ, R. et coll. (2019). Exploring transfer learning and domain data selection for the biomedical translation. Dans *Proceedings of the Fourth Conference on Machine Translation (Volume 3 : Shared Task Papers, Day 2)*, pages 156–163.

ADAMS, W. G., PHILLIPS, B. D., BACIC, J. D., WALSH, K. E., SHANAHAN, C. W. et PAASCHE-ORLOW, M. K. (2014). Automated conversation system before pediatric primary care visits : a randomized trial. *Pediatrics*, 134(3):e691–e699.

AGRAWAL, A. (2009). Medication errors : prevention using information technology systems. *British journal of clinical pharmacology*, 67(6):681–686.

AL-ZUBAIDE, H. et ISSA, A. A. (2011). Ontbot : Ontology based chatbot. Dans *International Symposium on Innovations in Information and Communications Technology*, pages 7–12. IEEE.

ALLEN, J. F. et PERRAULT, C. R. (1980). Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.

ALSENTZER, E., MURPHY, J. R., BOAG, W., WENG, W.-H., JIN, D., NAUMANN, T. et MCDERMOTT, M. (2019). Publicly available clinical bert embeddings. *arXiv preprint arXiv :1904.03323*.

ALTIERI, R., INCARDONA, F., KIRKILIS, H. et RICCI, R. (2006). Mobi-dev : Mobile devices for healthcare applications. Dans *M-Health : Emerging Mobile Health Systems*, pages 163–175. Springer US, Boston, MA.

AMIN, S., DUNFIELD, K. A., VECHKAEVA, A. et NEUMANN, G. (2020). A data-driven approach for noise reduction in distantly supervised biomedical relation extraction. *arXiv preprint arXiv :2005.12565*.

AMIN-NEJAD, A., IVE, J. et VELUPILLAI, S. (2020). Exploring transformer text generation for medical dataset augmentation. Dans *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708.

AMINI, R., LISETTI, C., YASAVUR, U. et RISHE, N. (2013). On-demand virtual health counselor for delivering behavior-change health interventions. Dans *2013 IEEE International Conference on Healthcare Informatics*, pages 46–55. IEEE.

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

ANNASAMY, R. M. et SYCARA, K. (2019). Towards better interpretability in deep q-networks. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4561–4569.

ARAMAKI, E., KANO, Y., OHKUMA, T. et MORITA, M. (2016). Mednlpdoc : Japanese shared task for clinical nlp. Dans *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 13–16.

ARAMAKI, E., MASKAWA, S. et MORITA, M. (2011). Twitter catches the flu : detecting influenza epidemics using twitter. Dans *Proceedings of the 2011 Conference on empirical methods in natural language processing*, pages 1568–1576.

ARONSON, A. R. (2001). Effective mapping of biomedical text to the umls metathesaurus : the metamap program. Dans *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

ARONSON, A. R. et LANG, F.-M. (2010). An overview of metamap : historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

ASPDEN, P., WOLCOTT, J., BOOTMAN, J. L. et CRONENWETT, L. R. (2006). Committee on identifying and preventing medication errors : preventing medication errors. *Institute of Medicine National Academy Press, Washington, DC*.

AUGRY, F., RAVAUD, P., LOPEZ, I., LETELLIER, D., ILTIS, A., BOUSCARY, D. et HAZEBROUCQ, G. (1998). Erreurs de prescription des médicaments cytotoxiques : étude prospective de 5 000 ordonnances. *Journal de Pharmacie Clinique*, 17(1):20–4.

AUSTIN, J. L. (1975). *How to do things with words*. Oxford university press.

BABU, M., HEMCHANDHAR, R., AKASH, S., TODI, A. et coll. (2021). Voice prescription with end-to-end security enhancements. Dans *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pages 1–8. IEEE.

BACH, K. et HARNISH, R. M. (1979). *Linguistic Communication and Speech Acts*. Cambridge : MIT Press.

BAHDANAU, D., CHO, K. et BENGIO, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

BAPNA, A., TUR, G., HAKKANI-TUR, D. et HECK, L. (2017). Sequential dialogue context modeling for spoken language understanding. *arXiv preprint arXiv :1705.03455*.

BARKER, K. N., FLYNN, E. A., PEPPER, G. A., BATES, D. W. et MIKEAL, R. L. (2002). Medication errors observed in 36 health care facilities. *Archives of internal medicine*, 162(16):1897–1903.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. Dans *First international conference on language resources and evaluation (LREC)*, pages 1373–1376.

BELL, S., WOOD, C. et SARKAR, A. (2019). Perceptions of chatbots in therapy. Dans *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

BELTAGY, I., LO, K. et COHAN, A. (2019). Scibert : A pretrained language model for scientific text. *arXiv preprint arXiv :1903.10676*.

BÉRARD, A., PIETQUIN, O., SERVAN, C. et BESACIER, L. (2016). Listen and translate : A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv :1612.01744*.

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

BHATT, V., LI, J. et MAHARJAN, B. (2021). Docpal : A voice-based ehr assistant for health practitioners. Dans *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*, pages 1–6. IEEE.

BIBAULT, J.-E., CHAIX, B., GUILLEMASSÉ, A., COUSIN, S., ESCANDE, A., PERRIN, M., PIENKOWSKI, A., DELAMON, G., NECTOUX, P. et BROUARD, B. (2019). A chatbot versus physicians to provide information for patients with breast cancer : Blind, randomized controlled noninferiority trial. *Journal of medical Internet research*, 21(11):e15787.

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

BIRD, S. (2006). Nltk : the natural language toolkit. Dans *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

BISSOTO, A., VALLE, E. et AVILA, S. (2021). Gan-based data augmentation and anonymization for skin-lesion analysis : A critical review. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856.

BLACK, L.-A., MCTEAR, M., BLACK, N., HARPER, R. et LEMON, M. (2005). Appraisal of a conversational artefact and its utility in remote patient monitoring. Dans *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pages 506–508. IEEE.

BOBROW, D. G., KAPLAN, R. M., KAY, M., NORMAN, D. A., THOMPSON, H. et WINOGRAD, T. (1977). Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173.

BODENREIDER, O. (2004). The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

BORDES, A., BOUREAU, Y.-L. et WESTON, J. (2016). Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv :1605.07683*.

BRINKMAN, W.-P., HARTANTO, D., KANG, N., de VLIEGHER, D., KAMPMANN, I. L., MORINA, N., EMMELKAMP, P. G. et NEERINCX, M. (2012). A virtual reality dialogue system for the treatment of social phobia. Dans *CHI'12 extended abstracts on human factors in computing systems*, pages 1099–1102.

BROWN, E. G., WOOD, L. et WOOD, S. (1999). The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A. et coll. (2020). Language models are few-shot learners. *arXiv preprint arXiv :2005.14165*.

BUONSIGNORI, C. (2003). Les erreurs médicamenteuses et le circuit du médicament anticancéreux. Mémoire de D.E.A., Ecole Nationale de la Santé Publique.

BURGUN, A., NAMER, F., RUCH, P. et LE DUFF, F. (2003). Umlf : construction d'un lexique médical francophone unifié.

CAELEN, J. (2003). Stratégies de dialogue. Dans *Conférence MFI*, volume 3, pages 20–22.

CAMBRIA, E. et WHITE, B. (2014). Jumping nlp curves : A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

CAMPILLOS, L., DELÉGER, L., GROUIN, C., HAMON, T., LIGOZAT, A.-L. et NÉVÉOL, A. (2018). A french clinical corpus with comprehensive semantic annotations : development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601.

CAMPILLOS-LLANOS, L., THOMAS, C., BILINSKI, É., ZWEIGENBAUM, P. et ROSSET, S. (2020). Designing a virtual patient dialogue system based on terminology-rich resources : Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

CARCHIOLO, V., LONGHEU, A., REITANO, G. et ZAGARELLA, L. (2019). Medical prescription classification : a nlp-based approach. Dans *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 605–609. IEEE.

CARPENTER, R. (2007). Jabberwacky.

CHALKIDIS, I., FERGADIOTIS, M., MALAKASIOTIS, P., ALETRAS, N. et ANDROUTSOPOULOS, I. (2020). Legal-bert : The muppets straight out of law school. *arXiv preprint arXiv :2010.02559*.

CHAN, W., JAITLY, N., LE, Q. et VINYALS, O. (2016). Listen, attend and spell : A neural network for large vocabulary conversational speech recognition. Dans *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE.

CHAPMAN, W. W., NADKARNI, P. M., HIRSCHMAN, L., D'AVOLIO, L. W., SAVOVA, G. K. et UZUNER, O. (2011). Overcoming barriers to nlp for clinical text : the role of shared tasks and the need for additional creative solutions.

CHARNOCK, R. (1999). Les langues de spécialité et le langage technique : considérations didactiques. *ASp. la revue du GERAS*, (23-26):281–302.

CHELBA, C., MAHAJAN, M. et ACERO, A. (2003). Speech utterance classification. Dans *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

CHEN, H., LIU, X., YIN, D. et TANG, J. (2017). A survey on dialogue systems : Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

CHEN, Y.-N., CELIKYILMAZ, A. et HAKKANI-TUR, D. (2018). Deep learning for dialogue systems. Dans *Proceedings of the 27th International Conference on Computational Linguistics : Tutorial Abstracts*, pages 25–31.

CHIU, J. P. et NICHOLS, E. (2016). Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

CHOROWSKI, J., BAHDANAU, D., CHO, K. et BENGIO, Y. (2014). End-to-end continuous speech recognition using attention-based recurrent nn : First results. *arXiv preprint arXiv :1412.1602*.

CHU-CARROLL, J. et BROWN, M. K. (1997). Tracking initiative in collaborative dialogue interactions. *arXiv preprint cmp-lg/9704005*.

CLAVEAU, V. et ZWEIGENBAUM, P. (2005). Traduction de termes biomédicaux par inférence de transducteurs. Dans *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*, pages 251–260.

COLBY, K. M., WEBER, S. et HILF, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, 2(1):1–25.

COLLINS, F. S. et TABAK, L. A. (2014). Policy : Nih plans to enhance reproducibility. *Nature News*, 505(7485):612.

COSSIN, S., LOUSTAU, R., JOUHET, V., LÉTINIER, L., MOUGIN, F., EVRARD, G., GIL-JARDINÉ, C., DIALLO, G. et THIESSARD, F. (2018). Romedi, une terminologie médicale française pour la détection des médicaments en texte libre. *Artificial Intelligence Platform, Nancy*.

CÔTÉ, R. A., ROTHWELL, D. J. et BROCHU, L. (1993). Snomed international. *Northfield, IL : College of American Pathologists*.

de BRUIJN, B., CHERRY, C., KIRITCHENKO, S., MARTIN, J. et ZHU, X. (2010). Nrc at i2b2 : one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. Dans *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA : i2b2*.

DE BRUIJN, B., CHERRY, C., KIRITCHENKO, S., MARTIN, J. et ZHU, X. (2011). Machine-learned solutions for three stages of clinical information extraction : the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

DELCROIX, M.-H. et GOMEZ, C. (2020). *Ordonnances en gynécologie obstétrique : 103 prescriptions courantes*. Maloine.

DEMNER-FUSHMAN, D., CHAPMAN, W. W. et MCDONALD, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

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

DENNY, J. C., IRANI, P. R., WEHBE, F. H., SMITHERS, J. D. et SPICKARD III, A. (2003). The knowledgemap project : development of a concept-based medical school curriculum database. Dans *AMIA Annual Symposium Proceedings*, volume 2003, page 195. American Medical Informatics Association.

DERIU, J., RODRIGO, A., OTEGI, A., ECHEGOYEN, G., ROSSET, S., AGIRRE, E. et CIELIEBAK, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

DESOT, T., PORTET, F. et VACHER, M. (2019a). Slu for voice command in smart home : comparison of pipeline and end-to-end approaches. Dans *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 822–829. IEEE.

DESOT, T., PORTET, F. et VACHER, M. (2019b). Towards end-to-end spoken intent recognition in smart home. Dans *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

DESOT, T., RAIMONDO, S., MISHAKOVA, A., PORTET, F. et VACHER, M. (2018). Towards a French Smart-Home Voice Command Corpus : Design and NLU Experiments. Dans *21st International Conference on Text, Speech and Dialogue TSD 2018*, Brno, Czech Republic.

DEVLIN, J., CHANG, M.-W., LEE, K. et TOUTANOVA, K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

DHINGRA, B., LI, L., LI, X., GAO, J., CHEN, Y.-N., AHMED, F. et DENG, L. (2016). Towards end-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv :1609.00777*.

DINARELLI, M. et ROSSET, S. (2011). Models cascade for tree-structured named entity detection. Dans *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1269–1278.

DLIGACH, D., MILLER, T. et SAVOVA, G. K. (2015). Semi-supervised learning for phenotyping tasks. Dans *AMIA annual symposium proceedings*, volume 2015, page 502. American Medical Informatics Association.

DOAN, S., BASTARACHE, L., KLIMKOWSKI, S., DENNY, J. et XU, H. (2009). Vanderbilt's system for medication extraction. Dans *Proceedings of the Third i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*.

DOS SANTOS, M. C., MONTYNE, F. et DHAEN, C. (2006). Medical natural language processing enhancing drug ordering and coding. Dans ISTEPANIAN, R. S. H., LAXMINARAYAN, S. et PATTICHIS, C. S., éditeurs : *M-Health : Emerging Mobile Health Systems*, pages 147–161. Springer, Boston, MA.

DOUGHERTY, R. C. (2013). *Natural language computing : An English generative grammar in prolog*. Psychology Press.

DUNLOP, A. J. (2014). *Efficient Latent-Variable Grammars : Learning and Inference*. Thèse de doctorat, Citeseer.

DYBKJÆR, L., BERNSEN, N. O. et DYBKJÆR, H. (1996). Grice incorporated : cooperativity in spoken dialogue. Dans *COLING 1996 Volume 1 : The 16th International Conference on Computational Linguistics*.

DZIKOVSKA, M. O., MOORE, J. D., STEINHAUSER, N. et CAMPBELL, G. (2011). Exploring user satisfaction in a tutorial dialogue system.

EL BOUKKOURI, H. (2020). Ré-entraîner ou entraîner soi-même? stratégies de pré-entraînement de bert en domaine médical. Dans *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 29–42. ATALA ; AFCP.

ERIC, M. et MANNING, C. D. (2017). A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *arXiv preprint arXiv :1701.04024*.

ETHIER, J.-F., BARTON, A. et TASEEN, R. (2018). An ontological analysis of drug prescriptions. *Applied Ontology*, 13(4):273–294.

ETHIER, J.-F., TASEEN, R., LAVOIE, L. et BARTON, A. (2016). Improving the semantics of drug prescriptions with a realist ontology. Dans *ICBO/BioCreative*.

FEDUS, W., GOODFELLOW, I. et DAI, A. M. (2018). Maskgan : better text generation via filling in the_. *arXiv preprint arXiv :1801.07736*.

FILLMORE, C. J. et coll. (1976). Frame semantics and the nature of language. Dans *Annals of the New York Academy of Sciences : Conference on the origin and development of language and speech*, volume 280, pages 20–32. New York.

FILLMORE, C. J. et BAKER, C. F. (2001). Frame semantics for text understanding. Dans *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, volume 6.

FITZPATRICK, K. K., DARCY, A. et VIERHILE, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot) : a randomized controlled trial. *JMIR mental health*, 4(2):e19.

FOUQUET, Y. (2004). *Modélisation des attentes en dialogue oral*. Thèse de doctorat, Université Joseph-Fourier-Grenoble I.

FRASER, K. C., NEJADGHOLI, I., DE BRUIJN, B., LI, M., LAPLANTE, A. et ABIDINE, K. Z. E. (2019). Extracting umls concepts from medical text using general and domain-specific deep learning models. *arXiv preprint arXiv :1910.01274*.

FRASER, N. M. et GILBERT, G. N. (1991). Simulating speech systems. *Computer Speech & Language*, 5(1):81–99.

FRIEDMAN, C. (1997). Towards a comprehensive medical language processing system : methods and issues. Dans *Proceedings of the AMIA annual fall symposium*, page 595. American Medical Informatics Association.

FRIEDMAN, C. (2000). A broad-coverage natural language processing system. Dans *Proceedings of the AMIA Symposium*, page 270. American Medical Informatics Association.

GAFFNEY, H., MANSELL, W. et TAI, S. (2019). Conversational agents in the treatment of mental health problems : Mixed-method systematic review. *JMIR Mental Health*, 6(10):e14166.

GAL, A., LAPALME, G. et SOMERS, H. (1991). *Prolog for natural language processing*. Wiley Chichester.

GAŠIĆ, M. et YOUNG, S. (2013). Gaussian processes for pomdp-based dialogue manager optimization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):28–40.

GEHRING, J., AULI, M., GRANGIER, D., YARATS, D. et DAUPHIN, Y. N. (2017). Convolutional sequence to sequence learning. Dans *ICML 2017*, page 1243–1252.

GEHRMANN, S., DAI, F. Z., ELDER, H. et RUSH, A. M. (2018). End-to-end content and plan selection for data-to-text generation. *arXiv preprint arXiv :1810.04700*.

GIORGINO, T., AZZINI, I., ROGNONI, C., QUAGLINI, S., STEFANELLI, M., GRETTER, R. et FALAVIGNA, D. (2005). Automated spoken dialogue system for hypertensive patient home management. *International Journal of Medical Informatics*, 74(2-4):159–167.

GLIGIC, L., KORMILITZIN, A., GOLDBERG, P. et NEVADO-HOLGADO, A. (2020). Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. *Neural Networks*, 121:132–139.

GOLOVNEVA, O. et PERIS, C. (2020). Generative adversarial networks for annotated data augmentation in data sparse nlu. *arXiv preprint arXiv :2012.05302*.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. et BENGIO, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

GOURIEUX, B. (Octobre 2019). Le circuit du médicamenten établissement de santé - rôle de la pharmacie clinique. Présentation Universitaire.

GRICE, H. P. (1957). Meaning. *The philosophical review*, 66(3):377–388.

GRICE, H. P. (1975). Logic and conversation. Dans *Speech acts*, pages 41–58. Brill.

GRICE, H. P., COLE, P. et MORGAN, J. L. (1975). Syntax and semantics.

GRIESSHABER, D., MAUCHER, J. et VU, N. T. (2020). Fine-tuning bert for low-resource natural language understanding via active learning. *arXiv preprint arXiv :2012.02462*.

GUNASEKARA, C., KIM, S., D'HARO, L. F., RASTOGI, A., CHEN, Y.-N., ERIC, M., HEDAYATNIA, B., GOPALAKRISHNAN, K., LIU, Y., HUANG, C.-W. et coll. (2020). Overview of the ninth dialog system technology challenge : Dstc9. *arXiv preprint arXiv :2011.06486*.

GUZMAN, B., METZGER, I., APHINYANAPHONGS, Y., GROVER, H. et coll. (2020). Assessment of amazon comprehend medical : Medication information extraction. *arXiv preprint arXiv :2002.00481*.

HAFFNER, P., TUR, G. et WRIGHT, J. H. (2003). Optimizing svms for complex call classification. Dans *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

HAURY, B. et CASES, C. (2005). Les événements indésirables graves liés aux soins observés dans les établissements de santé : premiers résultats d'une étude nationale.

HECK, M., van NIEKERK, C., LUBIS, N., GEISHAUSER, C., LIN, H.-C., MORESI, M. et GAŠIĆ, M. (2020). Trippy : A triple copy strategy for value independent neural dialog state tracking. *arXiv preprint arXiv :2005.02877*.

HEMPHILL, C. T., GODFREY, J. J. et DODDINGTON, G. R. (1990). The atis spoken language systems pilot corpus. Dans *Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

HENDERSON, J., LEMON, O. et GEORGILA, K. (2005). Hybrid reinforcement/supervised learning for dialogue policies from communicator data. Dans *IJCAI workshop on knowledge and reasoning in practical dialogue systems*.

HENDERSON, J., LEMON, O. et GEORGILA, K. (2008). Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4):487–511.

HENDERSON, M., GAŠIĆ, M., THOMSON, B., TSIAKOULIS, P., YU, K. et YOUNG, S. (2012). Discriminative spoken language understanding using word confusion networks. Dans *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 176–181. IEEE.

HENDERSON, M., THOMSON, B. et YOUNG, S. (2014). Word-based dialog state tracking with recurrent neural networks. Dans *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299.

HÖGBERG, J. (1997). Data driven formant synthesis. Dans *Fifth European Conference on Speech Communication and Technology*.

HOWARD, J. et RUDER, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv :1801.06146*.

HOY, M. B. (2018). Alexa, siri, cortana, and more : an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88.

HUANG, J., OSORIO, C. et SY, L. W. (2019). An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177:141–153.

HUANG, L., SIL, A., JI, H. et FLORIAN, R. (2017). Improving slot filling performance with attentive neural networks on dependency structures. *arXiv preprint arXiv :1707.01075*.

HUANG, Z., XU, W. et YU, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv :1508.01991*.

HUDLICKA, E. (2013). Virtual training and coaching of health behavior : Example from mindfulness meditation training. *Patient education and counseling*, 92(2):160–166.

HUTCHENS, J. L. (1996). How to pass the turing test by cheating. *School of Electrical, Electronic and Computer Engineering research report TR97-05. Perth : University of Western Australia*.

IKHU-OMOREGBE, N. et AZETA, A. (2010). A voice-based mobile prescription application for healthcare services (vbmopa). *IJECS-IJENS*, 10(2):73–78.

JAGANNATHA, A. N. et YU, H. (2016). Bidirectional rnn for medical event detection in electronic health records. Dans *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2016, page 473. NIH Public Access.

JAVED, F., BRYANT, B. R., ČREPINŠEK, M., MERNIK, M. et SPRAGUE, A. (2004). Context-free grammar induction using genetic programming. Dans *Proceedings of the 42nd annual Southeast regional conference*, pages 404–405.

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

JEONG, M. et LEE, G. G. (2009). Multi-domain spoken language understanding with transfer learning. *Speech Communication*, 51(5):412–424.

JIANG, M., CHEN, Y., LIU, M., ROSENBLOOM, S. T., MANI, S., DENNY, J. C. et XU, H. (2011). A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606.

JIANG, S. et de RIJKE, M. (2018). Why are sequence-to-sequence models so dull? *EMNLP 2018*, page 81.

JIN, Q., DHINGRA, B., COHEN, W. et LU, X. (2019). Probing biomedical embeddings from language models. Dans *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89.

JOHNSON, A., BULGARELLI, L., POLLARD, T., HORNG, S., CELI, L. A. et MARK IV, R. (2020). Mimic-iv (version 0.4). *PhysioNet*.

JOHNSON, A. E., POLLARD, T. J., SHEN, L., LI-WEI, H. L., FENG, M., GHASSEMI, M., MOODY, B., SZOLOVITS, P., CELI, L. A. et MARK, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

JURAFSKY, D. (2000). *Speech & language processing*. Pearson Education India.

JURAFSKY, D. et MARTIN, J. H. (2014). Speech and language processing. vol. 3. *US : Prentice Hall*.

JWALAPURAM, P. (2017). Evaluating dialogs based on grice's maxims. Dans *Proceedings of the Student Research Workshop associated with RANLP*, pages 17–24.

KALLEL, R. (1999). La créativité lexicale dans la langue de la médecine. *Initial (e) s*, 18:106–127.

KALYAN, K. S., RAJASEKHARAN, A. et SANGEETHA, S. (2021). Ammu–a survey of transformer-based biomedical pretrained language models. *arXiv preprint arXiv :2105.00827*.

KAUSHAL, R., BATES, D. W., LANDRIGAN, C., MCKENNA, K. J., CLAPP, M. D., FEDERICO, F. et GOLDMANN, D. A. (2001). Medication errors and adverse drug events in pediatric inpatients. *Jama*, 285(16):2114–2120.

KEARNS, W. R., CHI, N.-C., CHOI, Y. K., LIN, S.-Y., THOMPSON, H. et DEMIRIS, G. (2019). A systematic review of health dialog systems. *Methods of information in medicine*, 58(06): 179–193.

KELLY, L., SUOMINEN, H., GOEURIOT, L., NEVES, M., KANOULAS, E., LI, D., AZZOPARDI, L., SPIJKER, R., ZUCCON, G., SCELLS, H. et coll. (2019). Overview of the clef ehealth evaluation lab 2019. Dans *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 322–339. Springer.

KIM, J.-w., KIM, S.-t., YOON, J.-y. et JOO, Y.-I. (2015). A personal prescription management system employing optical character recognition technique. *Journal of the Korea Institute of Information and Communication Engineering*, 19(10):2423–2428.

KOCABALLI, A. B., BERKOVSKY, S., QUIROZ, J. C., LARANJO, L., TONG, H. L., REZAZADEGAN, D., BRIATORE, A. et COIERA, E. (2019). The personalization of conversational agents in health care : Systematic review. *Journal of medical Internet research*, 21(11):e15360.

KOCABIYIKOGLU, A. C., BABOUCHKINE, J.-M., QADER, R. et PORTET, F. (2021). Neural medication extraction : A comparison of recent models in supervised and semi-supervised learning settings. Dans *ICHI 2021 : IEEE International Conference on Healthcare Informatics*.

KOCABIYIKOGLU, A. C., PORTET, F., BABOUCHKINE, J.-M. et BLANCHON, H. (2020). Spoken medical prescription acquisition through a dialogue system on smartphone : Perspective of a healthcare software company. Dans *LREC 2020 Industry Track Language Resources and Evaluation Conference 11–16 May 2020*.

KONKOL, M. et KONOPÍK, M. (2015). Segment representations in named entity recognition. Dans *International Conference on Text, Speech, and Dialogue*, pages 61–70. Springer.

KREIMEYER, K., FOSTER, M., PANDEY, A., ARYA, N., HALFORD, G., JONES, S. F., FORSHEE, R., WALDERHAUG, M. et BOTSIS, T. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information : a systematic review. *Journal of biomedical informatics*, 73:14–29.

KUNDIG, F. (2011). Médicamentslook-alike. *Rev Med Suisse*, 7:1955–61.

KURATA, G., XIANG, B., ZHOU, B. et YU, M. (2016). Leveraging sentence-level information with encoder lstm for semantic slot filling. *arXiv preprint arXiv :1601.01530*.

KUSNER, M. J. et HERNÁNDEZ-LOBATO, J. M. (2016). Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv preprint arXiv :1611.04051*.

LACHÈVRE, B. (2016). Les erreurs de délivrance à l'officine : état des lieux, prévention et gestion.

LAMBERT, B. L., GALANTER, W., LIU, K. L., FALCK, S., SCHIFF, G., RASH-FOANIO, C., SCHMIDT, K., SHRESTHA, N., VAIDA, A. J. et GAUNT, M. J. (2019). Automated detection of wrong-drug prescribing errors. *BMJ quality & safety*, 28(11):908–915.

LAMEL, L., ROSSET, S., GAUVAIN, J.-L. et BENNACEF, S. (1999). The limsi arise system for train travel information. Dans *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 501–504. IEEE.

LARANJO, L., DUNN, A. G., TONG, H. L., KOCABALLI, A. B., CHEN, J., BASHIR, R., SURIAN, D., GALLEGO, B., MAGRABI, F., LAU, A. Y. et coll. (2018). Conversational agents in healthcare : a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

LARIVEN, D. S. (2008). *Le Guide des premières ordonnances*. Editions de Santé.

LAU, F., KUZIEMSKY, C., PRICE, M. et GARDNER, J. (2010). A review on systematic reviews of health information system studies. *JAMIA*, 17(6):637–645.

LE, H., VIAL, L., FREJ, J., SEGONNE, V., COAVOUX, M., LECOUTEUX, B., ALLAUZEN, A., CRABBÉ, B., BESACIER, L. et SCHWAB, D. (2019). Flaubert : Unsupervised language model pre-training for french. *arXiv preprint arXiv :1912.05372*.

LE, H., VIAL, L., FREJ, J., SEGONNE, V., COAVOUX, M., LECOUTEUX, B., ALLAUZEN, A., CRABBÉ, B., BESACIER, L. et SCHWAB, D. (2020). Flaubert : des modèles de langue contextualisés pré-entraînés pour le français. Dans *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, pages 268–278. ATALA ; AFCP.

LE, H. T. (2007). A frame-based approach to text generation. Dans *Proceedings of the 21st Pacific Asia conference on language, information and computation*, pages 192–201.

LEE, J., SCOTT, D. J., VILLARROEL, M., CLIFFORD, G. D., SAEED, M. et MARK, R. G. (2011). Open-access mimic-ii database for intensive care research. Dans *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 8315–8318. IEEE.

LEE, J., YOON, W., KIM, S., KIM, D., KIM, S., SO, C. H. et KANG, J. (2020). Biobert : a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

LEE, K.-S. et COX, R. V. (2002). A segmental speech coder based on a concatenative tts. *Speech communication*, 38(1-2):89–100.

LEE, W., PARK, H., YOON, J., KIM, K. et CHOI, J. (2019). Clinical text generation through leveraging medical concept and relations. *arXiv preprint arXiv :1910.00861*.

LEMON, O., BRACY, A., GRUENSTEIN, A. et PETERS, S. (2001). The witas multi-modal dialogue system i. Dans *Seventh European Conference on Speech Communication and Technology*.

LEVIN, E. et PIERACCINI, R. (1995). Concept-based spontaneous speech understanding system. Dans *Fourth European Conference on Speech Communication and Technology*.

LEVIN, E., PIERACCINI, R. et ECKERT, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on speech and audio processing*, 8(1):11–23.

LI, J., SUN, A., HAN, J. et LI, C. (2020a). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

LI, X., CHEN, Y.-N., LI, L., GAO, J. et CELIKYILMAZ, A. (2017). End-to-end task-completion neural dialogue systems. *arXiv preprint arXiv :1703.01008*.

LI, X., FENG, J., MENG, Y., HAN, Q., WU, F. et LI, J. (2019). A unified mrc framework for named entity recognition. *arXiv preprint arXiv :1910.11476*.

LI, Z., KISELEVA, J. et de RIJKE, M. (2020b). Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. *arXiv preprint arXiv :2009.09781*.

LI, Z., LIU, F., ANTIEAU, L., CAO, Y. et YU, H. (2010). Lancet : a high precision medication event extraction system for clinical text. *Journal of the American Medical Informatics Association*, 17(5):563–567.

LIPSCOMB, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

LITMAN, D. et SILLIMAN, S. (2004). Itspoke : An intelligent tutoring spoken dialogue system. Dans *Demonstration papers at HLT-NAACL 2004*, pages 5–8.

LIU, B. et LANE, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv :1609.01454*.

LIU, B. et LANE, I. (2017). An end-to-end trainable neural network model with belief tracking for task-oriented dialog. *arXiv preprint arXiv :1708.05956*.

LIU, B. et LANE, I. (2018). End-to-end learning of task-oriented dialogs. Dans *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Student Research Workshop*, pages 67–73.

LIU, B., TUR, G., HAKKANI-TUR, D., SHAH, P. et HECK, L. (2018). Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. *arXiv preprint arXiv :1804.06512*.

LIU, F. et PEREZ, J. (2017). Gated end-to-end memory networks. Dans *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, pages 1–10.

LIU, F., TUR, G., HAKKANI-TÜR, D. et YU, H. (2011). Towards spoken clinical-question answering : evaluating and adapting automatic speech-recognition systems for spoken clinical questions. *Journal of the American Medical Informatics Association*, 18(5):625–630.

LIU, S., MA, W., MOORE, R., GANESAN, V. et NELSON, S. (2005). Rxnorm : prescription for electronic drug information exchange. *IT professional*, 7(5):17–23.

LIU, S., TANG, B., CHEN, Q. et WANG, X. (2015). Drug name recognition : approaches and resources. *Information*, 6(4):790–810.

LIU, W. et CAI, S. (2015). Translating electronic health record notes from english to spanish : A preliminary study. Dans *Proceedings of BioNLP 15*, pages 134–140.

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L. et STOYANOV, V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

LIYANAGE, H., KRAUSE, P. et de LUSIGNAN, S. (2015). Using ontologies to improve semantic interoperability in health data. *BMJ Health & Care Informatics*, 22(2):309–315.

LLANOS, L. C., BOUAMOR, D., BILINSKI, E., LIGOZAT, A.-L., ZWEIGENBAUM, P. et ROSSET, S. (2015). Description of the patientgenesys dialogue system. Dans *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 438–440.

LÓPEZ-CÓZAR, R., GARCÍA, P., DIAZ-VERDEJO, J. et RUBIO, A. J. (1997). A voice activated dialogue system for fast-food restaurant applications. Dans *Fifth European Conference on Speech Communication and Technology*.

LUCAS, B. (2000). Voicexml for web-based distributed conversational applications. *Communications of the ACM*, 43(9):53–57.

LUGOSCH, L., RAVANELLI, M., IGNOTO, P., TOMAR, V. S. et BENGIO, Y. (2019). Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv :1904.03670*.

LUONG, T., PHAM, H. et MANNING, C. D. (2015). Effective approaches to attention-based neural machine translation. Dans *EMNLP 2015*, pages 1412–1421.

MAHATPURE, J., MOTWANI, M. et SHUKLA, P. K. (2019). An electronic prescription system powered by speech recognition natural language processing and blockchain technology. *Int. J. Sci. Technol. Res.*, 8(8):1454–1462.

MARTIN, L., MULLER, B., SUÁREZ, P. J. O., DUPONT, Y., ROMARY, L., de LA CLERGERIE, É. V., SEDDAH, D. et SAGOT, B. (2019). Camembert : a tasty french language model. *arXiv preprint arXiv :1911.03894*.

MAURER, P. M. (1990). Generating test data with enhanced context-free grammars. *Ieee Software*, 7(4):50–55.

MCCRAY, A. T., SRINIVASAN, S. et BROWNE, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. Dans *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 235. American Medical Informatics Association.

MERABTI, T. (2010). *Méthodes pour la mise en relations des terminologies médicales : contribution a l'interopérabilité sémantique Inter et Intra terminologique*. Thèse de doctorat, Rouen.

MESNIL, G., DAUPHIN, Y., YAO, K., BENGIO, Y., DENG, L., HAKKANI-TUR, D., HE, X., HECK, L., TUR, G., YU, D. et coll. (2014). Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

MESNIL, G., HE, X., DENG, L. et BENGIO, Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. Dans *Interspeech*, pages 3771–3775.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S. et DEAN, J. (2013). Distributed representations of words and phrases and their compositionality. Dans *Advances in neural information processing systems*, pages 3111–3119.

MILLE, F., BOURDON, O., FONTAN, J. et BRION, F. (2005). Evaluation de la spécificité d'un système de détection automatisée des interactions médicamenteuses. *Acta Pharmacol Biol Clin*, 12:361–8.

MINER, A. S., MILSTEIN, A., SCHUELLER, S., HEGDE, R., MANGURIAN, C. et LINOS, E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5): 619–625.

MIRANDA-ESCALADA, A., GONZALEZ-AGIRRE, A., ARMENGOL-ESTAPÉ, J. et KRALLINGER, M. (2020). Overview of automatic clinical coding : annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. Dans *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

MISHAKOVA, A., PORTET, F., DESOT, T. et VACHER, M. (2019a). Learning natural language understanding systems from unaligned labels for voice command in smart homes. Dans *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 832–837. IEEE.

MISHAKOVA, A., PORTET, F., DESOT, T. et VACHER, M. (2019b). Learning Natural Language Understanding Systems from Unaligned Labels for Voice Command in Smart Homes. Dans *The 1st International Workshop on Pervasive Computing and Spoken Dialogue Systems Technology (PerDial 2019)*, Kyoto, Japan.

MORRIS, R. R., KOUDDOUS, K., KSHIRSAGAR, R. et SCHUELLER, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications : system design and user perceptions. *Journal of medical Internet research*, 20(6):e10148.

MÜLLER, M., SALATHÉ, M. et KUMMERVOLD, P. E. (2020). Covid-twitter-bert : A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv :2005.07503*.

NAKAYAMA, H., KUBO, T., KAMURA, J., TANIGUCHI, Y. et LIANG, X. (2018). doccano : Text annotation tool for human. Software available from https ://github.com/doccano/doccano.

NELSON, S. J., ZENG, K., KILBOURNE, J., POWELL, T. et MOORE, R. (2011). Normalized names for clinical drugs : Rxnorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448.

NÉVÉOL, A. (2018). *Traitement Automatique de la Langue Biomédicale*. Thèse de doctorat, Université Paris Sud.

NÉVÉOL, A., COHEN, K. B., GROUIN, C., HAMON, T., LAVERGNE, T., KELLY, L., GOEURIOT, L., REY, G., ROBERT, A., TANNIER, X. et coll. (2016). Clinical information extraction at the clef ehealth evaluation lab 2016. Dans *CEUR workshop proceedings*, volume 1609, page 28. NIH Public Access.

NÉVÉOL, A., DALIANIS, H., VELUPILLAI, S., SAVOVA, G. et ZWEIGENBAUM, P. (2018). Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of biomedical semantics*, 9(1):1–13.

NÉVÉOL, A., ZWEIGENBAUM, P., MAX, A., YVON, F., IVANISHCHEVA, Y. et RAVAUD, P. (2013). Statistical machine translation of systematic reviews into french. *Training*, 15(526):366K.

NING, Y., HE, S., WU, Z., XING, C. et ZHANG, L.-J. (2019). A review of deep learning based speech synthesis. *Applied Sciences*, 9(19):4050.

NÉVÉOL, A., GROUIN, C., LEIXA, J., ROSSET, S. et ZWEIGENBAUM, P. (2014). The QUAERO French medical corpus : A ressource for medical entity recognition and normalization. Dans *Proc of BioTextMining Work*, pages 24–30.

OMS, O. M. d. l. S. (1993). Cim-10 : Classification statistique internationale des maladies et des problèmes de santé connexes. Dans *CIM-10 : classification statistique internationale des maladies et des problèmes de santé connexes*, pages 1335–1335.

OORD, A. v. d., DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., KALCHBRENNER, N., SENIOR, A. et KAVUKCUOGLU, K. (2016). Wavenet : A generative model for raw audio. *arXiv preprint arXiv :1609.03499*.

OTT, M., EDUNOV, S., BAEVSKI, A., FAN, A., GROSS, S., NG, N., GRANGIER, D. et AULI, M. (2019). fairseq : A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv :1904.01038*.

PATRICK, J. et LI, M. (2009). A cascade approach to extracting medication events. Dans *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 99–103.

PATRICK, J. et LI, M. (2010). High accuracy information extraction of medication information from clinical notes : 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527.

PENG, B., LI, X., GAO, J., LIU, J., WONG, K.-F. et SU, S.-Y. (2018). Deep dyna-q : Integrating planning for task-completion dialogue policy learning. *arXiv preprint arXiv :1801.06176*.

PENG, B. et YAO, K. (2015). Recurrent neural networks with external memory for language understanding. *arXiv preprint arXiv :1506.00195*.

PENG, Y., YAN, S. et LU, Z. (2019). Transfer learning in biomedical natural language processing : an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv :1906.05474*.

PEREZ, J., BOUREAU, Y.-L. et BORDES, A. (2017). Dialog system & technology challenge 6 overview of track 1-end-to-end goal-oriented dialog learning. *Dialog System Technology Challenges*, 6.

PERROT, S. (2015). *Thérapeutique pratique 2015*. MED-LINE.

PHILIP, P., BIOULAC, S., SAUTERAUD, A., CHAUFTON, C. et OLIVE, J. (2014). Could a virtual human be used to explore excessive daytime sleepiness in patients ? *Presence : teleoperators and virtual environments*, 23(4):369–376.

PIAD-MORFFIS, A., GUTIÉRREZ, Y. et MUÑOZ, R. (2019). A corpus to support ehealth knowledge discovery technologies. *Journal of biomedical informatics*, 94:103172.

POVEY, D., GHOSHAL, A., BOULIANNE, G., BURGET, L., GLEMBEK, O., GOEL, N., HANNEMANN, M., MOTLICEK, P., QIAN, Y., SCHWARZ, P. et coll. (2011). The kaldi speech recognition toolkit. Dans *IEEE 2011 workshop on automatic speech recognition and understanding*, numéro CONF. IEEE Signal Processing Society.

POWELL, J. (2019). Trust me, i'ma chatbot : How artificial intelligence in health care fails the turing test. *Journal of medical Internet research*, 21(10):e16222.

PUZIKOV, Y. et GUREVYCH, I. (2018). E2e nlg challenge : Neural models vs. templates. Dans *Proceedings of the 11th International Conference on Natural Language Generation*, pages 463–471.

QADER, R., PORTET, F. et LABBÉ, C. (2019). Semi-supervised neural text generation by joint learning of natural language generation and natural language understanding models. *arXiv preprint arXiv :1910.03484*.

QADER, R., PORTET, F. et LABBÉ, C. (2020). Seq2seqpy : A lightweight and customizable toolkit for neural sequence-to-sequence modeling. Dans *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7140–7144.

QIU, Z., CHO, E., MA, X. et CAMPBELL, W. (2019). Graph-based semi-supervised learning for natural language understanding. Dans *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 151–158.

QUATTONI, A., WANG, S., MORENCY, L.-P., COLLINS, M. et DARRELL, T. (2007). Hidden conditional random fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(10):1848–1852.

QUDAR, M. M. A. et MAGO, V. (2020). Tweetbert : A pretrained language representation model for twitter text analysis. *arXiv preprint arXiv :2010.11091*.

RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., SUTSKEVER, I. et coll. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

RASTOGI, A., ZANG, X., SUNKARA, S., GUPTA, R. et KHAITAN, P. (2020). Towards scalable multi-domain conversational agents : The schema-guided dialogue dataset. Dans *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

RAYNER, M., BOUILLON, P., BROTANEK, J., FLORES, G., HALIMI, S., HOCKEY, B. A., ISAHARA, H., KANZAKI, K., KRON, E., NAKAO, Y. et coll. (2008). The 2008 medslt system. Dans *Coling 2008 : Proceedings of the workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, pages 32–35.

REBHOLZ-SCHUHMANN, D., CLEMATIDE, S., RINALDI, F., KAFKAS, S., van MULLIGEN, E. M., BUI, C., HELLRICH, J., LEWIN, I., MILWARD, D., POPRAT, M. et coll. (2013). Entity recognition in parallel multi-lingual biomedical corpora : The clef-er laboratory overview. Dans *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 353–367. Springer.

RUDNICKY, A. et XU, W. (1999). An agenda-based dialog management architecture for spoken language systems. Dans *IEEE Automatic Speech Recognition and Understanding Workshop*, volume 13.

SACKS, H. (1992). Lectures on conversation : Volume i. *Malden, Massachusetts : Blackwell*.

SANJEEV, S., PONNEKANTI, G. S. et REDDY, G. P. (2021). Advanced healthcare system using artificial intelligence. Dans *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 76–81. IEEE.

SARZYNSKI, E., DECKER, B., THUL, A., WEISMANTEL, D., MELARAGNI, R., CHOLAKIS, E., TEWARI, M., BECKHOLT, K., ZAROUKIAN, M., KENNEDY, A. C. et coll. (2017). Beta testing a novel smartphone application to improve medication adherence. *Telemedicine and e-Health*, 23(4):339–348.

SAVOVA, G. K., MASANZ, J. J., OGREN, P. V., ZHENG, J., SOHN, S., KIPPER-SCHULER, K. C. et CHUTE, C. G. (2010). Mayo clinical text analysis and knowledge extraction system (ctakes) : architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

SCHATZMANN, J., THOMSON, B., WEILHAMMER, K., YE, H. et YOUNG, S. (2007). Agenda-based user simulation for bootstrapping a pomdp dialogue system. Dans *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers*, pages 149–152.

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

SCHMITT, E. (2002). Iatrogénie médicamenteuse liée à la prescription en chimiothérapie anticancéreuse. *ONCOLOGIE-PARIS-*, 4(3):192–202.

SCHULZ, S., MARKÓ, K., SBRISSIA, E., NOHAMA, P. et HAHN, U. (2004). Cognate mapping-a heuristic strategy for the semi-supervised acquisition of a spanish lexicon from a portuguese seed lexicon. Dans *COLING 2004 : Proceedings of the 20th International Conference on Computational Linguistics*, pages 813–819.

SCULLEY, D., HOLT, G., GOLOVIN, D., DAVYDOV, E., PHILLIPS, T., EBNER, D., CHAUDHARY, V., YOUNG, M., CRESPO, J.-F. et DENNISON, D. (2015). Hidden technical debt in machine learning systems. Dans *Advances in neural information processing systems*, pages 2503–2511.

SEARLE, J. R. (1969). *Speech acts : An essay in the philosophy of language*.

SEE, A., LIU, P. J. et MANNING, C. D. (2017). Get to the point : Summarization with pointer-generator networks. *arXiv preprint arXiv :1704.04368*.

SENEFF, S. (1992). Tina : A natural language system for spoken language applications. *Computational linguistics*, 18(1):61–86.

SENNRICH, R., HADDOW, B. et BIRCH, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv :1508.07909*.

SERBAN, I. V., SORDONI, A., BENGIO, Y., COURVILLE, A. et PINEAU, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. Dans *Thirtieth AAAI Conference on Artificial Intelligence*.

SERPOLLET, N., BERGOUNIOUX, G., CHESNEAU, A. et WALTER, R. (2007). A large reference corpus for spoken French : Eslo 1 and 2 and its variations. Dans *Proceedings from Corpus Linguistics Conference Series, University of Birmingham*.

SHANG, L., LU, Z. et LI, H. (2015). Neural responding machine for short-text conversation. *arXiv preprint arXiv :1503.02364*.

SHEN, J., PANG, R., WEISS, R. J., SCHUSTER, M., JAITLY, N., YANG, Z., CHEN, Z., ZHANG, Y., WANG, Y., SKERRV-RYAN, R. et coll. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. Dans *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

SI, Y., WANG, J., XU, H. et ROBERTS, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11): 1297–1304.

SMITH, R. W. et GORDON, S. A. (1997). Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue. *Computational Linguistics*, 23(1): 141–168.

SPACKMAN, K. A., CAMPBELL, K. E. et CÔTÉ, R. A. (1997). Snomed rt : a reference terminology for health care. Dans *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

STALNAKER, R. C. (1978). Assertion. Dans *Pragmatics*, pages 315–332. Brill.

STEEDMAN, M. (1996). A very short introduction to ccg. *Unpublished paper. http ://www. coqsci. ed. ac. uk/steedman/paper. html*.

SU, P.-H., BUDZIANOWSKI, P., ULTES, S., GASIC, M. et YOUNG, S. (2017). Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. *arXiv preprint arXiv :1707.00130*.

SU, P.-H., GASIC, M., MRKSIC, N., ROJAS-BARAHONA, L., ULTES, S., VANDYKE, D., WEN, T.-H. et YOUNG, S. (2016a). Continuously learning neural dialogue management. *arXiv preprint arXiv :1606.02689*.

SU, P.-H., GASIC, M., MRKSIC, N., ROJAS-BARAHONA, L., ULTES, S., VANDYKE, D., WEN, T.-H. et YOUNG, S. (2016b). On-line active reward learning for policy optimisation in spoken dialogue systems. *arXiv preprint arXiv :1605.07669*.

SUKHBAATAR, S., SZLAM, A., WESTON, J. et FERGUS, R. (2015). End-to-end memory networks. *arXiv preprint arXiv :1503.08895*.

SUN, W., RUMSHISKY, A. et UZUNER, O. (2013). Evaluating temporal relations in clinical text : 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

SUTSKEVER, I., VINYALS, O. et LE, Q. V. (2014). Sequence to sequence learning with neural networks. Dans *Advances in neural information processing systems*, pages 3104–3112.

SUTTON, C., MCCALLUM, A. et ROHANIMANESH, K. (2007). Dynamic conditional random fields : Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8(3).

SYBRANDT, J. et SAFRO, I. (2020). Cbag : Conditional biomedical abstract generation. *arXiv preprint arXiv :2002.05637*.

TALBOT, T. B., KALISCH, N., CHRISTOFFERSEN, K., LUCAS, G. M. et FORBELL, E. (2016). Natural language understanding performance & use considerations in virtual medical encounters. Dans *MMVR*, pages 407–413.

TANG, B., WU, Y., JIANG, M., CHEN, Y., DENNY, J. C. et XU, H. (2013). A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835.

TAO, C., FILANNINO, M. et UZUNER, Ö. (2017). Prescription extraction using crfs and word embeddings. *Journal of biomedical informatics*, 72:60–66.

TAO, C., FILANNINO, M. et UZUNER, Ö. (2018). Fable : A semi-supervised prescription information extraction system. Dans *AMIA Annual Symposium proceedings*, volume 2018, page 1534. American Medical Informatics Association.

THOMPSON, C. A. (2008). Usp says thousands of drug names look or sound alike.

TOURILLE, J., FERRET, O., TANNIER, X. et NÉVÉOL, A. (2017). Temporal information extraction from clinical text. Dans *Conference of the European Chapter of the Association for Computational Linguistics*.

TRAUM, D. R. et HINKELMAN, E. A. (1992). Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599.

TUBAY, B. et COSTA-JUSSA, M. R. (2018). Neural machine translation with the transformer and multi-source romance languages for the biomedical wmt 2018 task. Dans *Proceedings of the Third Conference on Machine Translation : Shared Task Papers*, pages 667–670.

TUR, G. et DE MORI, R. (2011). *Spoken language understanding : Systems for extracting semantic information from speech*. John Wiley & Sons.

UZUNER, O., BODNARI, A., SHEN, S., FORBUSH, T., PESTIAN, J. et SOUTH, B. R. (2012). Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

UZUNER, Ö., SOLTI, I. et CADAG, E. (2010a). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.

UZUNER, Ö., SOLTI, I., XIA, F. et CADAG, E. (2010b). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association*, 17(5):519–523.

UZUNER, O., SZOLOVITS, P. et KOHANE, I. (2006). i2b2 workshop on natural language processing challenges for clinical records. Dans *Proceedings of the Fall Symposium of the American Medical Informatics Association*. Citeseer.

VAIDYAM, A. N., WISNIEWSKI, H., HALAMKA, J. D., KASHAVAN, M. S. et TOROUS, J. B. (2019). Chatbots and conversational agents in mental health : a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

VAN GURP, M., DECOENE, M., HOLVOET, M. et dos SANTOS, M. C. (2006). Linkbase, a philosophically-inspired ontology for nlp/nlu applications. Dans *KR-MED*.

VASWANI, A., BENGIO, S., BREVDO, E., CHOLLET, F., GOMEZ, A. N., GOUWS, S., JONES, L., KAISER, Ł., KALCHBRENNER, N., PARMAR, N. et coll. (2018). Tensor2tensor for neural machine translation. *arXiv preprint arXiv :1803.07416*.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. et POLOSUKHIN, I. (2017). Attention is all you need. Dans *Advances in neural information processing systems*, pages 5998–6008.

VECCHIATO, S. et GEROLIMICH, S. (2013). La langue médicale est-elle «trop complexe»? *Nouvelles perspectives en sciences sociales : revue internationale de systémique complexe et d'études relationnelles*, 9(1):81–122.

VELARDI, P., STILO, G., TOZZI, A. E. et GESUALDO, F. (2014). Twitter mining for fine-grained syndromic surveillance. *Artificial intelligence in medicine*, 61(3):153–163.

VINCENT, D. (2001). Les enjeux de l'analyse conversationnelle ou les enjeux de la conversation. *Revue québécoise de linguistique*, 30(1):177–198.

VINYALS, O. et LE, Q. (2015). A neural conversational model. *arXiv preprint arXiv :1506.05869*.

VLASOV, V., MOSIG, J. E. et NICHOL, A. (2019). Dialogue transformers. *arXiv preprint arXiv :1910.00486*.

VU, N. T. (2016). Sequential convolutional neural networks for slot filling in spoken language understanding. *arXiv preprint arXiv :1606.07783*.

WALKER, M., KAMM, C. et LITMAN, D. (2000). Towards developing general models of usability with paradise. *Natural Language Engineering*, 6(3-4):363–377.

WALKER, M. et WHITTAKER, S. (1990). Mixed initiative in dialogue : An investigation into discourse segmentation. Dans *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, pages 70–78. Association for Computational Linguistics.

WALLACE, R. S. (2009). The anatomy of alice. Dans *Parsing the Turing Test*, pages 181–210. Springer.

WANG, S., REN, F. et LU, H. (2018). A review of the application of natural language processing in clinical medicine. Dans *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 2725–2730. IEEE.

WANG, Y., DENG, L. et ACERO, A. (2011). Semantic frame-based spoken language understanding. *Spoken language understanding : systems for extracting semantic information from speech*, pages 41–91.

WANG, Y.-Y. et ACERO, A. (2002). Evaluation of spoken language grammar learning in the atis domain. Dans *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–41. IEEE.

WANG1, Y.-Y., ACERO, A., CHELBA, C., FREY, B. et WONG, L. (2002). Combination of statistical and rule-based approaches for spoken language understanding. Dans *Seventh International Conference on Spoken Language Processing*.

WARD, W. (1990). The cmu air travel information service : Understanding spontaneous speech. Dans *Speech and Natural Language : Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

WEIZENBAUM, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

WEN, T.-H., GASIC, M., KIM, D., MRKSIC, N., SU, P.-H., VANDYKE, D. et YOUNG, S. (2015a). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. *arXiv preprint arXiv :1508.01755*.

WEN, T.-H., GASIC, M., MRKSIC, N., SU, P.-H., VANDYKE, D. et YOUNG, S. (2015b). Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv :1508.01745*.

WEN, T.-H., VANDYKE, D., MRKSIC, N., GASIC, M., ROJAS-BARAHONA, L. M., SU, P.-H., ULTES, S. et YOUNG, S. (2016). A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv :1604.04562*.

WESTON, J., CHOPRA, S. et BORDES, A. (2014). Memory networks. *arXiv preprint arXiv :1410.3916*.

WILLIAMS, J., RAUX, A. et HENDERSON, M. (2016). The dialog state tracking challenge series : A review. *Dialogue & Discourse*, 7(3):4–33.

WILLIAMS, J. D., ASADI, K. et ZWEIG, G. (2017). Hybrid code networks : practical and efficient end-to-end dialog control with supervised and reinforcement learning. *arXiv preprint arXiv :1702.03274*.

WILLIAMS, J. D., POUPART, P. et YOUNG, S. (2008). Partially observable markov decision processes with continuous observations for dialogue management. Dans *Recent Trends in Discourse and Dialogue*, pages 191–217. Springer.

WILLIAMS, J. D. et ZWEIG, G. (2016). End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv :1606.01269*.

WONG, Y. W. et MOONEY, R. (2007). Generation by inverting a semantic parser that uses statistical machine translation. Dans *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*, pages 172–179.

WU, C.-S., MADOTTO, A., HOSSEINI-ASL, E., XIONG, C., SOCHER, R. et FUNG, P. (2019). Transferable multi-domain state generator for task-oriented dialogue systems. *arXiv preprint arXiv :1905.08743*.

WU, S., ROBERTS, K., DATTA, S., DU, J., JI, Z., SI, Y., SONI, S., WANG, Q., WEI, Q., XIANG, Y. et coll. (2020). Deep learning in clinical natural language processing : a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

WU, Y., JIANG, M., XU, J., ZHI, D. et XU, H. (2017). Clinical named entity recognition using deep learning models. Dans *AMIA Annual Symposium Proceedings*, volume 2017, page 1812. American Medical Informatics Association.

XU, H., STENNER, S. P., DOAN, S., JOHNSON, K. B., WAITMAN, L. R. et DENNY, J. C. (2010). Medex : a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24.

XU, P. et HU, Q. (2018). An end-to-end approach for handling unknown slot values in dialogue state tracking. *arXiv preprint arXiv :1805.01555*.

XU, P. et SARIKAYA, R. (2013). Convolutional neural network based triangular crf for joint intent detection and slot filling. Dans *2013 ieee workshop on automatic speech recognition and understanding*, pages 78–83. IEEE.

XU, W. et RUDNICKY, A. (2000). Task-based dialog management using an agenda. Dans *ANLP-NAACL 2000 Workshop : Conversational Systems*.

YAN, H., DENG, B., LI, X. et QIU, X. (2019). Tener : adapting transformer encoder for named entity recognition. *arXiv preprint arXiv :1911.04474*.

YANG, X., CHEN, Y.-N., HAKKANI-TÜR, D., CROOK, P., LI, X., GAO, J. et DENG, L. (2017). End-to-end joint learning of natural language understanding and dialogue manager. Dans *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5690–5694. IEEE.

YAO, K., PENG, B., ZHANG, Y., YU, D., ZWEIG, G. et SHI, Y. (2014a). Spoken language understanding using long short-term memory neural networks. Dans *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 189–194. IEEE.

YAO, K., PENG, B., ZWEIG, G., YU, D., LI, X. et GAO, F. (2014b). Recurrent conditional random field for language understanding. Dans *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4077–4081. IEEE.

YAO, K., ZWEIG, G., HWANG, M.-Y., SHI, Y. et YU, D. (2013). Recurrent neural networks for language understanding. Dans *Interspeech*, pages 2524–2528.

YAZDAVAR, A. H., AL-OLIMAT, H. S., EBRAHIMI, M., BAJAJ, G., BANERJEE, T., THIRUNARAYAN, K., PATHAK, J. et SHETH, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. Dans *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198.

YIN, P., LU, Z., LI, H. et KAO, B. (2015). Neural enquirer : Learning to query tables with natural language. *arXiv preprint arXiv :1512.00965*.

YOUNG, S., BRESLIN, C., GAŠIĆ, M., HENDERSON, M., KIM, D., SZUMMER, M., THOMSON, B., TSIAKOULIS, P. et HANCOCK, E. T. (2016). Evaluation of statistical pomdp-based dialogue systems in noisy environments. Dans *Situated Dialog in Speech-Based Human-Computer Interaction*, pages 3–14. Springer.

YOUNG, S., GAŠIĆ, M., KEIZER, S., MAIRESSE, F., SCHATZMANN, J., THOMSON, B. et YU, K. (2010). The hidden information state model : A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

YOUNG, S. J. (2000). Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences*, 358(1769):1389–1402.

YOUNG, T., HAZARIKA, D., PORIA, S. et CAMBRIA, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

ZHANG, J.-G., HASHIMOTO, K., WU, C.-S., WAN, Y., YU, P. S., SOCHER, R. et XIONG, C. (2019a). Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *arXiv preprint arXiv :1910.03544*.

ZHANG, Y., CHEN, Q., YANG, Z., LIN, H. et LU, Z. (2019b). Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.

ZHANG, Y. et LU, Z. (2019). Exploring semi-supervised variational autoencoders for biomedical relation extraction. *Methods*, 166:112–119.

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

ZHU, J., XIA, Y., WU, L., HE, D., QIN, T., ZHOU, W., LI, H. et LIU, T.-Y. (2020). Incorporating bert into neural machine translation. *arXiv preprint arXiv :2002.06823*.

ZHU, X. et GOLDBERG, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

ZWEIGENBAUM, P. (1997). Traitement automatique de la langue médicale.

ZWEIGENBAUM, P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, 2(5).

ZWEIGENBAUM, P., BACHIMONT, B., BOUAUD, J., CHARLET, J. et BOISVIEUX, J.-F. (1995). Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods of information in medicine*, 34(01/02):15–24.

ZWEIGENBAUM, P., DARMONI, S. J. et GRABAR, N. (2001). The contribution of morphological knowledge to french mesh mapping for information retrieval. Dans *Proceedings of the AMIA Symposium*, page 796. American Medical Informatics Association.

Bibliographie personnelle

[1] Ali Can Kocabiyikoglu et al. "Neural Medication Extraction: A Comparison of Recent Models in Supervised and Semi-supervised Learning Settings". In: *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*. IEEE. 2021, pp. 148–152.

[2] Ali Can Kocabiyikoglu et al. "Spoken Medical Prescription Acquisition Through a Dialogue System on Smartphone: Perspective of a Healthcare Software Company". In: *LREC 2020 Industry Track Language Resources and Evaluation Conference 11–16 May 2020*. 2020.

[3] Ali Can Kocabiyikoglu et al. "Towards spoken medical prescription understanding". In: *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE. 2019, pp. 1–8.

ANNEXE A

Caractéristiques des prescriptions médicamenteuses

Lorsqu'on évoque « une prescription », on pense souvent à la prescription d'un médicament. Même si la prescription de médicaments constituent la grande majorité de cas de prescription, un LAP doit prendre en compte toutes les catégories de prescription, telles que :

- Les traitements médicamenteux : destinés aux pharmacies pour un usage ambulatoire ou hospitalier
- Les préparations médicamenteuses : destinées aux pharmacies
- Les perfusions : souvent destiné à un usage hospitalier mais peut également être à domicile sous surveillance d'un professionnel de santé
- Les transfusions : à usage hospitalier
- Les demandes d'assistance respiratoire : principalement à usage hospitalier mais peut également être à domicile
- Les demandes d'analyses biologiques : destinées aux laboratoires d'analyses
- Les demandes de surveillance et de soins : destinées aux patients, à d'autres spécialistes (p. ex. psychiatre) ou aux personnels de santé (p. ex. infirmiers)
- Les demandes d'examens et de consultation : destinées aux spécialistes et d'autres professionnels de santé (ex. ostéopathie)
- Les demandes de rééducation : destinées aux masseurs-kinésithérapeutes
- Des consignes : destinées aux patients
- Des dispositifs médicaux : destinés aux patients ou pharmacies

Sur les ordonnances papiers, les prescriptions étant destinées à des acteurs différents, il est d'usage de rédiger une ordonnance par spécialiste concerné. Par exemple, lors d'une consultation, le patient peut recevoir plusieurs ordonnances : par exemple une pour les médicaments et une autre pour un examen biologique. Cependant, différentes types de prescriptions peuvent apparaître sur la même ordonnance s'ils sont destinées au même spécialiste. Par exemple, un médecin peut ajouter un conseil sur la prise d'un rendez-vous dans X temps après avoir suite à la prescription d'un ou plusieurs médicaments. Les caractéristiques de chaque catégorie de prescription sont détaillées dans les paragraphes suivants.

Les traitements médicamenteux

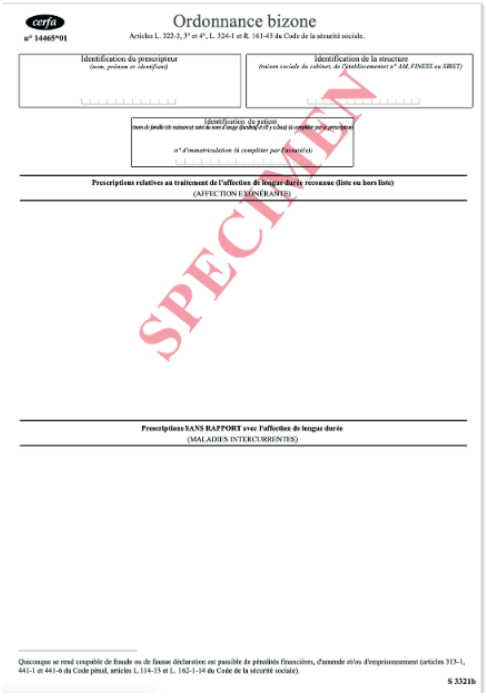
Les traitements médicamenteux constituent la grande majorité de cas de prescription. Comme la prescription médicamenteuse englobe différents types de traitements, il existe des sous-catégories spécifiques comme ce qui suit :

— Ordonnance simple : ordonnance comprenant des prescriptions médicamenteuses habituelles. Sur une ordonnance simple, il peut y avoir la prescription d'un produit (ex. savon avec un ph neutre).

— Ordonnance « bizone » : Une ordonnance dite bizone concerne les patients suivis pour une maladie chronique (le diabète, l'hypertension artérielle, l'asthme, etc.) La Figure A.1 montre un formulaire *CERFA* dédié à cette ordonnance spécifique. Sur la Figure A.1, la partie du haut est réservée à la maladie chronique (qui est pris en charge à 100% par la sécurité sociale). La partie du bas est dédiée à la prescription de médicaments hors du cadre de la maladie chronique et comporte les mêmes caractéristiques qu'une ordonnance standard.

— Ordonnance de médicaments, de produits ou de prestations d'exception : Certains produits ou médicaments sont soumis à des formalités particulières. Dans cette catégorie se trouvent notamment les médicaments à prescription initiale hospitalière réservée à certains spécialistes. Il concerne également les prescriptions de médicaments particulièrement onéreux.

— Ordonnance de médicaments stupéfiants : La prescription d'un médicament stupéfiant est réservée aux médecins, chirurgiens-dentistes et aux sages-femmes. Contrairement à une ordonnance habituelle, afin d'éviter les potentielles tentatives de modification, le prescripteur doit prescrire en toutes lettres, le nombre d'unités thérapeutiques par prise. Ce type d'ordonnance est également limité en durée : 7, 14 ou 28 jours.


cerfa
n° 14065*01

Ordonnance bizone

Articles L. 322-3, 3° et 4°, L. 324-1 et R. 161-45 du Code de la sécurité sociale.

Identification du prescripteur
(nom, prénom et identifiant)

Identification de la structure
(raison sociale du cabinet, du l'établissement n° EM, FINESS ou SIRET)

Identification du patient
(nom de famille (et naissance) suivi du nom d'usage (facultatif s'il y a lieu) à compléter par le prescripteur)

n° d'immatriculation (à compléter par l'assuré(e))

Prescriptions relatives au traitement de l'affection de longue durée reconnue (liste ou hors liste)
(AFFECTION EXONERANTE)

SPECIMEN

Prescriptions SANS RAPPORT avec l'affection de longue durée
(MALADIES INTERCURRENTES)

Quiconque se rend coupable de fraude ou de fausse déclaration est passible de pénalités financières, d'amende et/ou d'emprisonnement (articles 313-1, 441-1 et 441-6 du Code pénal, articles L.114-13 et L. 162-1-14 du Code de la sécurité sociale).

S 3321b

FIGURE A.1 – Exemple d'une ordonnance bizone

Même si les caractéristiques ci-dessus concernent la rédaction des ordonnances papiers, ils concernent également les prescriptions électroniques. Par exemple, les prescripteurs sont obligés d'ajouter dans certains cas d'ajouter un commentaire en expliquant le contexte particulier de la prescription. Un exemple simple de prescription médicamenteuse est donné ci-dessous.

Paracétamol 500 mg 2 à 6 comprimés par jour en fonction des douleurs pendant 7 jours

Les préparations médicamenteuses

La préparation médicamenteuse concerne les médicaments préparés en dehors des laboratoires pharmaceutiques. Comme une ordonnance standard, cette ordonnance est destinée aux pharmaciens. En revanche, au niveau sémantique, ce type de prescription peut-être très différent d'une prescription médicamenteuse standard. Il concerne les ingrédients et leurs dosages à mélanger pour constituer un produit final comme dans une recette de cuisine. Un exemple de prescription d'une préparation médicamenteuse est comme suivant :

Diprosone 30 gr, Cérat de Galien 70 gr, 1 pot de 100 g à appliquer 1 à 2 fois par jour sur le visage

Les perfusions La prescription des perfusions est similaire à celle des médicaments. Toutefois, sur ce type de prescription, il y a des indications plus complexes liée aux durées d'administration ainsi qu'aux instructions spécifiques destinées aux spécialistes. La prescription de perfusion suivante illustre cette complexité :

Métronidazole 500 mg en perfusion de 30 minutes dans 125 cc de SGI x3/j

Les transfusions

Les prescriptions concernant les transfusions (de sang ou d'autres produits) sont réservées à usage hospitalier et peuvent être très techniques. La sémantique employée dans une prescription médicamenteuse et une transfusion est de nature donc très différente. L'exemple de prescription suivant tiré de Lariven (2008) montre un exemple de prescription de transfusion.

Transfusion de X Culots Globulaires isoGR iso Rh , 1g=1CG=3 points d'hématocrite (1 PFC/3 CG)

Les demandes d'assistance respiratoire

Comme pour les transfusions, la prescription d'une oxygénothérapie ou d'une assistance respiratoire est très technique. La prescription doit contenir le type de ventilation, le type de ventilateur, les réglages spécifiques et la durée du traitement. Comme pour les transfusions, cette catégorie est très différente en termes d'information par rapport aux prescriptions médicamenteuses.

Les demandes d'analyses biologiques

Les demandes d'examens biologiques constituent l'une des catégories de prescription la plus courantes après les prescriptions médicamenteuses. Il existe trois catégories d'examens : les tests unitaires, les analyses et les bilans. Les bilans comportent souvent une série

d'analyses typiques liées à un diagnostic et sont composés de tests et d'examens. En termes d'information, il y a une riche nomenclature de noms de tests, parfois en langue étrangère qui constitue une difficulté importante. Souvent, les prescripteurs énumèrent les noms des tests à effectuer et ajoutent des remarques sur des tests spécifiques. L'exemple ci-dessous montre une prescription de demande d'analyse biologique typique.

Bilan para-clinique : NFS, urée, glycémie, calcémie, ionogramme, créatinimie, hémoculture, ...

Les demandes de surveillance et de soins

Ce type de prescription est souvent destiné à un infirmier (IDE) permettant de surveiller l'état du patient notamment suite à la sortie d'hospitalisation. Un exemple d'une liste de signes de gravité à surveiller pendant l'épidémie de Covid-19 a été comme ce qui suit :

Polypnée (fréquence respiratoire > 22/min), pression artérielle systolique < 90 mmHg, ...

Les demandes d'examens et consultations

Cette catégorie de prescription concerne la prescription d'un examen ou d'une consultation avec un spécialiste (ex. dermatologue, diététicien, etc.) parfois décrivant de façon succinct l'avis du médecin qui prescrit. En général, c'est le médecin traitant dans le cadre du parcours de soins qui effectue la demande d'une consultation qui permet une prévention personnalisée. Les demandes d'examens d'imagerie médicale sont souvent encadrées mais peuvent contenir des détails techniques. Un exemple de demande d'examen d'imagerie médicale est comme suivant :

Radiographie du bassin de face, faux profil de Lequesne des 2 hanches

Les demandes de rééducation

La rééducation concerne le patient qui suit un traitement visant une potentielle récupération de ses capacités antérieures suite à une maladie ou un accident. Les informations sur une ordonnance destinées à un kinésithérapeute concernent le type de rééducation et les indications particulières si nécessaire. Le prescripteur peut également préciser un nombre de séances sur la prescription. L'exemple tiré de Lariven (2008) montre une demande de rééducation :

Rééducation active et passive des membres déficitaires, rééducation de la marche et de l'équilibre

Des consignes

Les prescriptions sont souvent accompagnées de consignes. Pour un patient souffrant d'une maladie du coeur, cela peut-être une consigne simple comme « éviter le sel ». Les consignes ne font pas partie d'une ordonnance à part, ils ont plutôt un rôle accompagnateur. En termes d'information, les consignes sont très libres et contiennent de façon succincte des instructions souvent hygiéno-diététiques. Un exemple tiré de Lariven (2008) montre un conseil prescrit dans le cadre de l'hypertension artérielle du sujet âgé :

- Le matin au réveil ou dans la nuit ne jamais vous lever d'un seul coup mais en deux fois, en commençant par vous asseoir au bord du lit, ceci afin d'éviter les malaises

Prescription des dispositif médicaux

Les dispositifs médicaux sont une catégorie d'instruments ou de matériels plus grand et varié tels que des lunettes, implantes, des stents, etc. Comme pour les consignes, souvent ils font partie d'une ordonnance standard. Les prescriptions des dispositifs médicaux peuvent être très différentes selon la nature du dispositif médical prescrit. Selon les consignes d'utilisation du dispositif médical, une prescription peut contenir beaucoup d'instructions détaillées. L'exemple tiré de Lariven (2008) montre une ligne d'ordonnance prescrite pour un ulcère variqueux.

- Compression avec une pression comprise entre 30 et 40 mm Hg à la cheville, adaptée au cas par cas, en utilisant un bandage multicouche réalisé par la superposition de bandes à étirement courts (MEDICA©) et/ou long (BIFLEX©) ou par le dispositif PROFORE©.

A.0.1 Conclusion sur les caractéristiques des prescriptions médicamenteuses

Dans cette section, nous avons présenté en détail les différents types de prescriptions avec un exemple de prescription pour certains. Ces prescriptions comportent des caractéristiques différentes des uns les autres qui sont importants à prendre en compte lors de la modélisation du domaine. Même s'il y a des catégories telles que les consignes qui peuvent être très libres en termes de forme, et d'informations, nous pourrons voir des généralités. Dans la généralité, on retrouve une forme d'interaction courte et concise qui s'effectue par le biais d'un ensemble d'instructions à suivre, souvent encadré par des intervalles de temps précis. On retrouve également une technicité plus élevée lorsque la prescription est adressée à un professionnel de santé (ex. transfusions)

Dans cette thèse, pour restreindre le domaine sémantique, nous nous sommes limités aux prescriptions médicamenteuses. Cela nous a permis de cerner le domaine et constituer des cadres représentant une sémantique générale qui contiennent des informations communes applicables dans d'autres domaines de prescription. Par exemple, le nom d'un produit ou d'un médicament, les conditions et mentions d'administration (ou d'application ou d'utilisation), le rythme et la fréquence des prises ou des rendez-vous, la durée totale d'utilisation ou de prise de médicament ou le nombre de séances totales prévues.

ANNEXE B

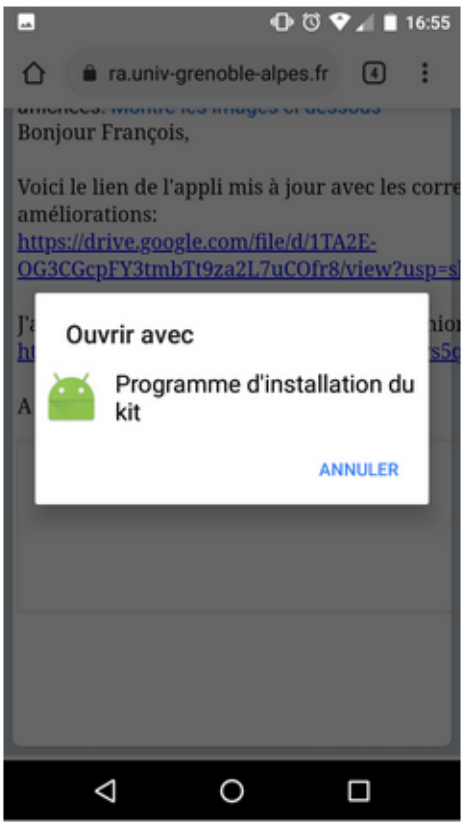
Procédure et guide d'installation de l'application sur smartphones Android

B.1 Introduction

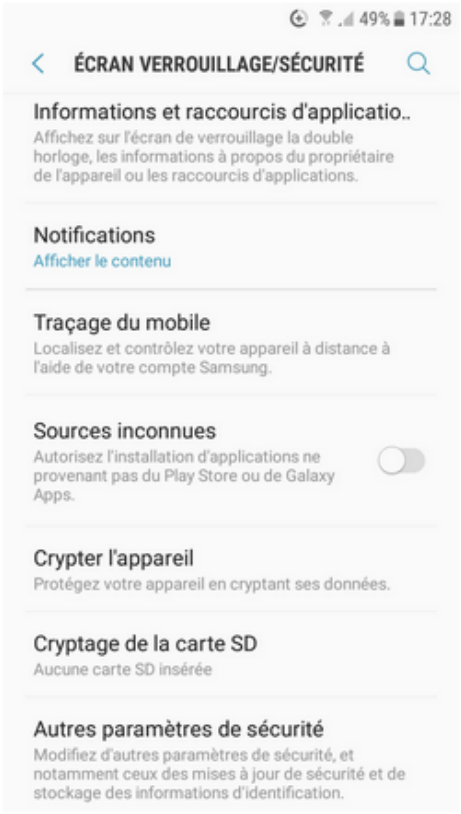
Étant donné l'épidémie de covid-19 et les mesures sanitaires en vigueur, l'ensemble de ces tests sont effectués de manière distance sur votre propre matériel ou un matériel prêté désinfecté. Ce document explique la procédure de test pas à pas de l'installation de l'application à tester jusqu'au téléchargement des données et le questionnaire d'enquête.

B.2 Installation de l'application

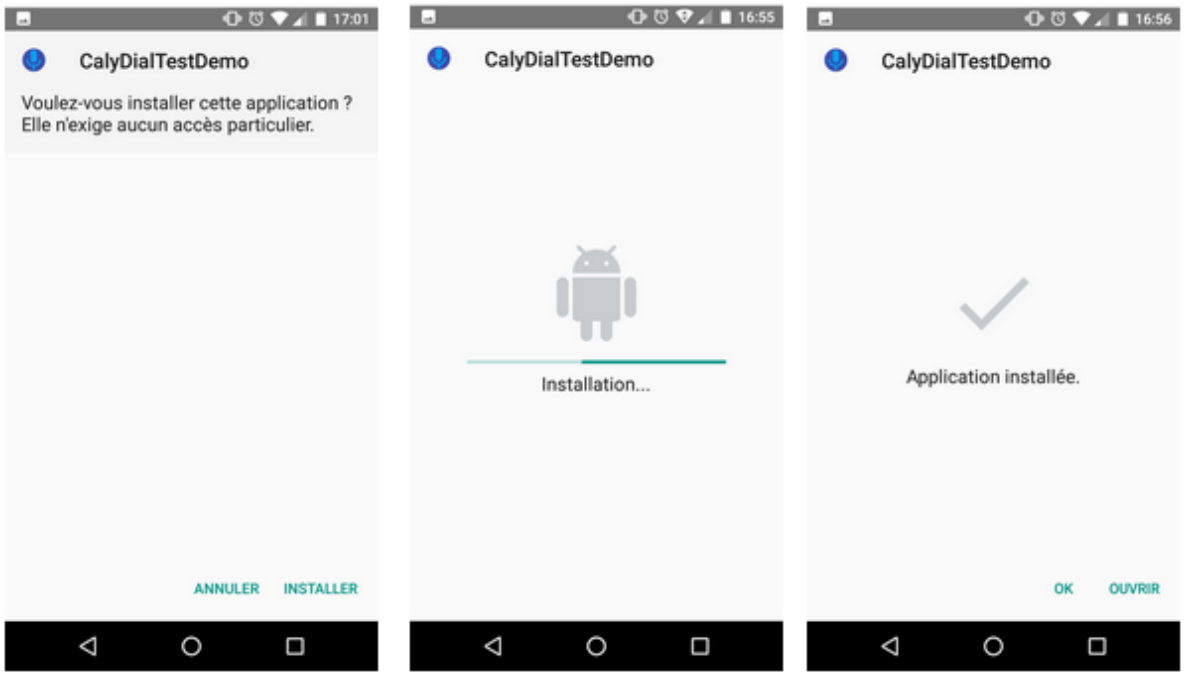
Étape 1 : Cliquez le lien qui vous a été envoyé par e-mail permettant d'installer l'application



Étape 1-bis : L'application d'un APK (application Android) pourrait nécessiter l'activation de sources inconnues dans certains portables. Si c'est le cas, activez cette option sur Paramètres > Sécurité & Confidentialité. Vous pouvez retirer cette option une fois l'application installée.

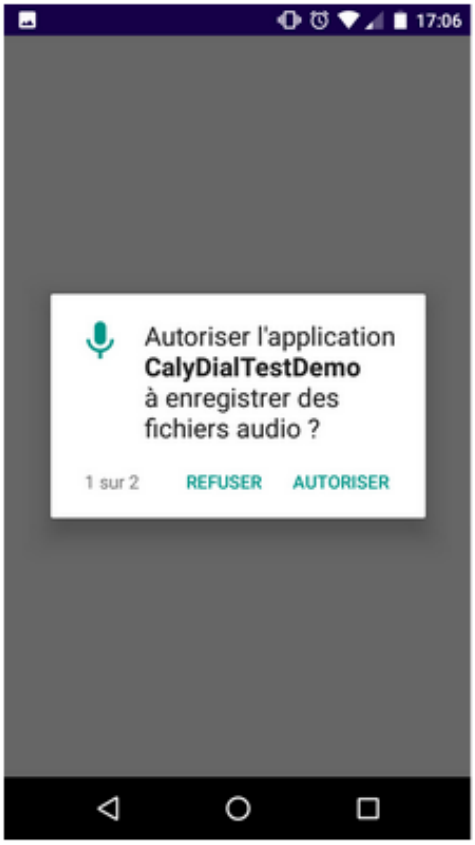


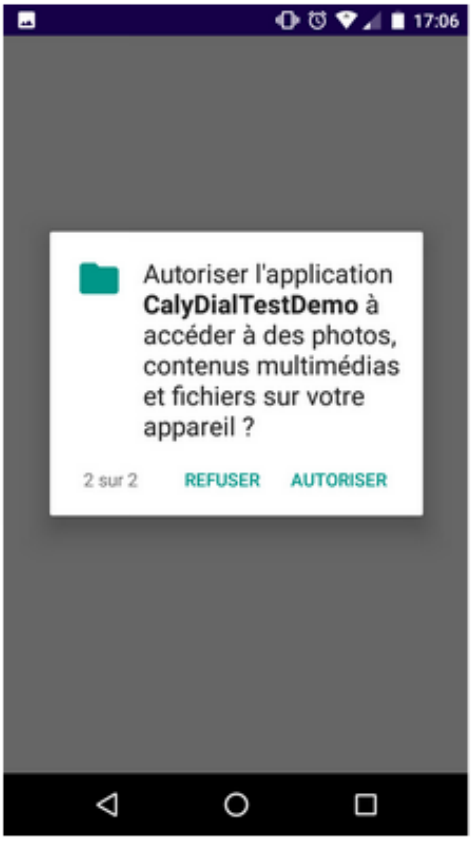
Étape 2 : Confirmez l'installation



Étape 3 : Ouvrez l'application puis donnez les autorisations nécessaires : la capture de

son et la sauvegarde des fichiers sur votre smartphone.



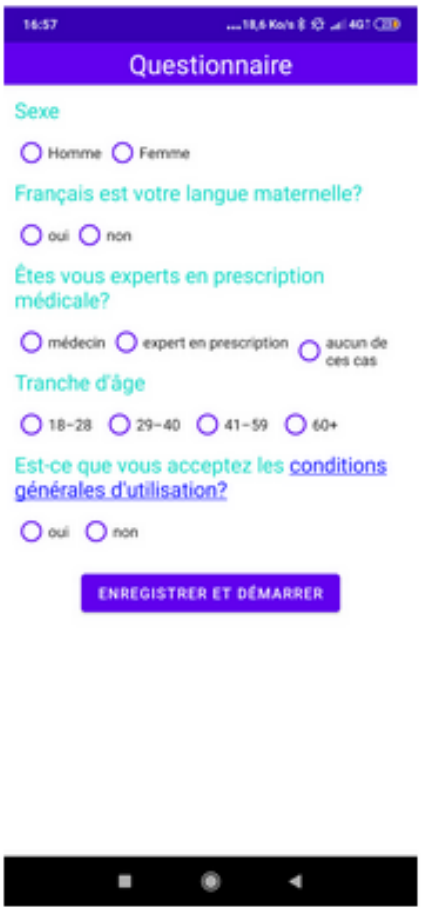


B.3 Phase d'expérimentation

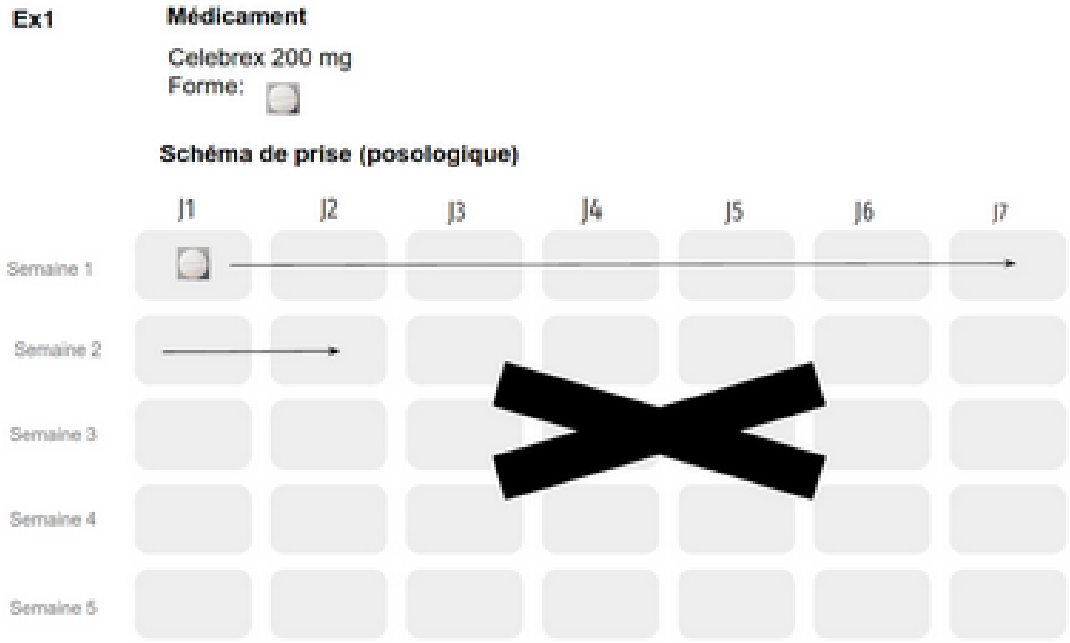
Étape 4 : Le questionnaire et les conditions générales d'utilisation

Ce questionnaire nous permet de prendre connaissance des données qui seront enregistrées et nous indique si vous souhaitez participer à l'expérimentation. Les conditions générales d'utilisation sont détaillées dans la dernière question. Vous y apprendrez que l'application n'enregistre pas d'autres données personnelles que votre voix et que l'application ne vous met pas en danger. Au cours de l'expérience, vous êtes libre d'arrêter à tout moment et de ne pas transmettre les données. Pour sortir de l'expérimentation, il suffit de désinstaller l'application et de nous avertir pour que nous supprimions les données.

Une fois que vous transmettez les données, il ne nous sera plus possible de les associer à votre nom. Elles ne pourront donc pas être supprimées.



Étape 5 : Munissez-vous des exemples de prescriptions graphiques tel que celui ci-dessous. Il est conseillé d'ouvrir les exemples qui vous sont fournis sur un autre appareil (ordinateur portable, tablette, etc.) pour faciliter le déroulement de l'expérience.



Pour le premier exemple, faites vous une représentation mentale de la prescription et

vocalisez là en dehors de l'application. Lorsque vous êtes prêt, passez à l'étape suivante.

Étape 5-bis : Munissez-vous des exemples de prescriptions à lire. Il est conseillé d'ouvrir les exemples qui vous sont fournis sur un autre appareil (ordinateur portable, tablette, etc.) pour faciliter le déroulement de l'expérience. Quelques exemples de réalisation :

1. Paracétamol 500mg, un comprimé par jour pendant une semaine
2. spasfon©, un comprimé par jour pendant une semaine, deux comprimés si douleur persistante

Faites vous une représentation mentale de la prescription et vocalisez là en dehors de l'application. Lorsque vous êtes prêt, passez à l'étape suivante.

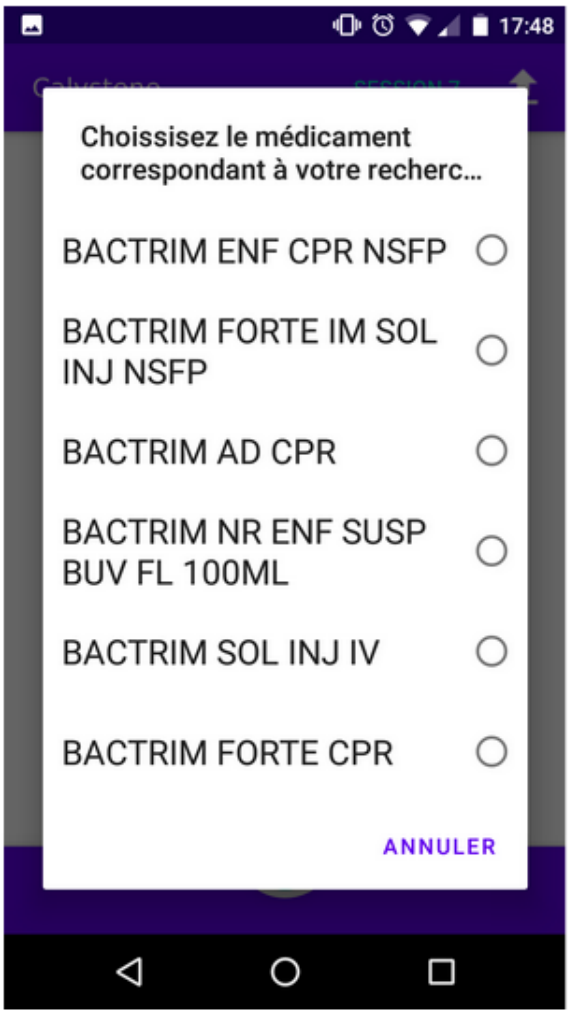
Étape 6 : Pour effectuer une prescription cliquez sur le bouton d'enregistrement. Vous pouvez effectuer votre enregistrement **après le bip sonore**. Puis lorsque vous avez terminé, cliquez sur le bouton d'arrêt situé au même endroit pour terminer.

1) Démarrez l'enregistrement

2) Après le bip, dites oralement votre prescription

3) arrêtez l'enregistrement

Étape 7 : Lorsque le système propose des choix de médicaments, vous pouvez choisir celui qui correspond à votre recherche avec votre doigt ou annuler pour réessayer. Si vous ne savez pas lequel prendre, prenez celui qui semble le plus proche. Ensuite laissez vous guider par le système.



Conseil : si le dialogue semble coincé, repliez-vous sur une prescription simple tel que paracétamol une dose par jour pendant une semaine. Et passez à la suite

Étape 8 : Quand la prescription est terminée, utilisez "ajouter un commentaire libre" pour répéter la prescription d'une traite. Vous pouvez également changer la posologie ou le médicament simplement en appuyant de nouveau sur le bouton. Vous pouvez également annuler la prescription si vous le désirez. Si la prescription est satisfaisante (même si elle ne reproduit pas fidèlement ce que vous avez dit) vous pouvez valider et passer à la suite.

Étape 9 : Munissez-vous des exemples de prescriptions à lire et suivez les mêmes étapes que précédemment pour les lire au système. Encore une fois, pour chaque exemple deux versions sont attendues : l'une par le dialogue, l'autre par l'ajout du commentaire.

Étape 10 : Quand vous avez fini l'ensemble des prescriptions vous pouvez télécharger vos données sur le serveur du lig via le bouton en haut à droite (Flèche dirigée vers le haut). Cela vous fait sortir de la session et vous (ou une autre personne) peut recommencer le processus.

B.4 Fin de l'expérimentation

Étape 11 : A la fin de l'expérience nous vous remercions de bien vouloir remplir un questionnaire qui permettra d'évaluer les points forts et points faibles de ce système à destination des professionnels de santé. Recevez tous nos remerciements pour votre implication et votre temps. Nous vous remercions de bien vouloir supprimer l'application de votre smartphone (paramètres -> Applications -> CalyDialTestDemo-> Désinstaller) .

Au bout de quelques semaines l'application ne sera plus opérationnelle. Il est donc inutile de la garder sur votre smartphone.

ANNEXE C

Convention de transcription orthographique

1. La transcription doit correspondre fidèlement au signal à l'exception des noms de médicaments et les dénominations commune internationales. Si le nom d'un médicament ou d'un molécule est mal prononcé transcrit le nom du médicament correct.
2. La transcription ne doit pas contenir d'erreurs d'orthographe ni de grammaire sur les mots, il faut respecter l'orthographe standard du français.
3. Ne pas tenir compte des élisions.
4. Ne pas corriger les erreurs grammaticales des participants.
5. Transcrire les nombres sous forme de chiffre sans espace entre les chiffres. L'usage des lettres permet d'écrire les grands nombres ronds.
6. Tout nombre décrivant les doses, dosages, posologies, durées de prescriptions, intervalles, rythmes et fréquences doivent être transcrits avec les chiffres.
7. Transcrire toutes les onomatopées et interjections produites par les locuteurs.
8. Transcrire les hésitations grâce à « euh » et « hum ».
9. Transcrire tous les mots entendus même s'ils sont répétés.
10. Marquer la troncation d'un mot par le symbole « / » en début ou fin de mot.
11. Ne pas corriger les lapsus et transcrire le mot tel qu'il a été prononcé.
12. Ne pas corriger les lapsus et transcrire le mot tel qu'il a été prononcé.
13. Transcrire le mot inconnu avec l'orthographe la plus proche de ce qui est entendu.