Rapport de stage

Annotations précises de dialogues pathologiques

Decker Amandine

Stage encadré par : Manuel Rebuschi et Maxime Amblard

Année universitaire 2019/2020

Coordonnées

IDMC

Site web : http://institut-sciences-digitales.fr
Pôle Herbert Simon, 13 rue Michel Ney, 54000 Nancy
idmc-contact@univ-lorraine.fr
+33 3 72 74 16 18

Archives Henri-Poincaré

Site web : https://poincare.univ-lorraine.fr
91 Avenue de la Libération, 54000 Nancy
archives-poincare-contact@univ-lorraine.fr
+33 3 72 74 15 84

Remerciements

Je tiens à remercier Manuel Rebuschi et Maxime Amblard pour m'avoir encadrée pendant ce stage. Leurs nombreux conseils et les discussions que nous avons eus m'ont permis d'avancer dans mon travail et de passer trois mois très enrichissants.

Je remercie également les membres de l'équipe Sémagramme pour m'avoir accueillie au sein leurs réunions d'équipe ainsi que les membres des Archives Henri Poincaré pour avoir consacré du temps à présenter leur travail durant une vidéoconférence dédiée aux stagiaires.

Tout ceci m'a permis de mieux comprendre le fonctionnement du monde de la recherche et j'en suis très reconnaissante.

Table des matières

1 Introduction

L'interaction conversationnelle est au centre des relations humaines. Elle est régie par certaines règles (Grice, 1975), tant sociales que cognitives, qui permettent l'intercompréhension. De ce fait, si au cours d'une discussion l'un des locuteurs s'affranchit de ces règles, on fait face à des difficultés qu'il est intéressant d'analyser pour essayer de préciser leurs caractéristiques d'un point de vue sémantique mais aussi pour tenter de proposer des solutions d'interprétation. De telles difficultés apparaissent plus manifestement lors d'échanges entre un locuteur ordinaire et une personne avec schizophrénie, c'est pourquoi étudier de telles conversations en les modélisant formellement peut aider à mieux comprendre et interpréter ces dysfonctionnements.

Le projet SLAM (Schizophrénie et langage : analyse et modélisation) (Amblard et al., 2012) s'inscrit dans un programme de recherche pluridisciplinaire mêlant des approches psycholinguistiques, pragmatico- et sémantico-formelles ainsi que philosophique. L'objectif est de mettre en évidence et d'analyser les discontinuités survenant dans des conversations dites *pathologiques* entre un psychologue et une personne avec schizophrénie afin d'aider au diagnostic de cette maladie.

Trois tâches complémentaires cohabitent au sein de ce projet. Il s'agit d'une part de recueillir des données pour constituer et étudier les corpus (Amblard et al., 2019b). D'autre part, il faut imaginer une représentation permettant de formaliser les échanges (Amblard et al., 2011) : l'objectif est d'utiliser un modèle inspiré de la logique pour l'appliquer aux conversations pathologiques et mettre en évidence les dysfonctionnements. Enfin, il faut chercher et mettre en œuvre des procédures permettant d'évaluer les méthodes et la modélisation employées.

Ce stage concerne l'adaptation et l'utilisation du modèle dans le cadre des conversations pathologiques. Jusqu'ici, il permet de repérer des discontinuités au niveau sémantico-pragmatique dans des extraits assez courts. Il s'agit en effet de relier chaque segment conversationnel de la discussion à un autre qui le précède en assignant une relation discursive à chaque liaison. Le but est de mettre en évidence le déroulement de l'échange et le rôle de chaque acte de dialogue par rapport à ce qui a déjà été dit. Les règles qui régissent l'utilisation de ces relations permettent de définir les discontinuités, les relations discursives sont donc primordiales dans ce modèle et le choix de l'ensemble de relations utilisables lors de l'annotation est important. Il faut en effet que cet ensemble soit suffisamment vaste pour que les annotations soient précises tout en restant facile à utiliser. Cet ensemble est jusqu'ici plutôt restreint, ce qui convient pour des extraits assez courts mais peut poser problèmes pour l'annotation de discussions plus complètes.

De plus, les travaux (Amblard et al., 2019a) et (Boritchev & Amblard, 2019) constituent une base pour préciser les annotations en questions-réponses du modèle d'annotation. Ces travaux ont une visée compositionnelle, c'est à dire que l'objectif est d'aboutir à un système automatique qui fait des calculs dans le cadre des questions et réponses, ce qui n'est pas le cas du projet SLAM où l'interaction est étudie de façon programmatique : c'est l'enchaînement des actes de dialogue qui est étudié de façon a obtenir une vision globale de la conversation. Mais s'inspirer de la précision qu'ils apportent aux interactions question-réponse peut permettre de mieux appréhender la contribution des questions dans les conversations pathologiques.

L'objectif de ce stage est d'étudier le rôle des questions et des réponses dans les conversations pathologique afin de préciser le modèle d'annotations puis de l'appliquer à des nouvelles transcriptions. J'ai donc commencé par annoter précisément les questions et réponses dans deux entretiens. Comme cette première étude ne mettait aucun dysfonctionnement en évidence, j'ai annoté tous les segments conversationnels avec un modèle d'annotation comportant plus de relations rhétoriques que pour le corpus précédent car les textes annotés étaient plus longs et présentaient donc des acte de dialogue avec des fonctions différentes que dans des extraits très courts. Ces annotations extensives ont mis en évidence un schéma de plus haut niveau lié à l'enchaînement des thèmes, ce qui a amené à réfléchir à une représentation supplémentaire, plus abstraite, pour compléter le premier modèle.

Ce stage s'est déroulé en télétravail avec des prises de contact avec l'équipe Sémagramme et des membres des Archives Henri Poincaré. Ces équipes mènent des projets de recherches pluridisciplinaires notamment liés à l'analyse et la modélisation du discours.

Dans ce rapport, après une présentation de Sémagramme et des Archives Henri Poincaré dans la partie 2, le travail bibliographique permettant la compréhension globale des objectif du projet SLAM et du fonctionnement de la modélisation choisie sera mis en avant dans la partie 3. Puis le travail sur le modèle d'annotation et l'analyse de nouvelles transcriptions seront expliqués dans la partie 4. Enfin, la partie 5 conclura sur le travail effectué tout au long du stage et exposera les perspectives ouvertes par celui-ci.

2 Présentation des organismes d'accueil

J'ai effectué mon stage sous la supervision de Manuel Rebuschi et Maxime Amblard. Le premier travaille aux Archives Henri Poincaré et le second au Loria dans l'équipe Sémagramme, j'ai donc eu un aperçu des deux structures. Compte tenu des circonstances sanitaires, j'ai effectué l'intégralité de mon stage en télétravail. Cependant, j'ai pu découvrir les Archives ainsi que l'équipe Sémagramme grâce à des vidéoconférences

Une vidéoconférence par semaine avec mes tuteurs nous permettait de faire un point sur le travail que j'avais effectué et de réfléchir à ce qu'il fallait faire ensuite, j'ai pris part aux réunions hebdomadaires de l'équipe Sémagramme au cours desquelles j'ai pu découvrir les axes de travail de plusieurs des membres, enfin j'ai assisté à une présentation des Archives Henri Poincaré et d'une partie des membres de l'équipe.

SLAM est un projet sur lequel collaborent trois laboratoires : le Loria, les Archives Henri Poincaré et l'Atilf (Analyse et Traitement Informatique de la Langue Française). Sa gestion a longtemps reposé sur la Maison des Sciences de l'Homme, une institution dédiée aux projets interdisciplinaires. C'est aujourd'hui dans le cadre du projet OLKi que se poursuivent les recherches dans la continuité de SLAM.

2.1 OLKi

Open Language and Knowledge for Citizens (OLKi) est un projet interdisciplinaire mené dans le cadre de l'Initiative Lorraine Université d'Excellence (LUE) qui cherche à

développer de nouveaux algorithmes d'apprentissage automatique dédiés à l'extraction de connaissance à partir des données langagières. Ce projet accorde une importance particulière à la transparence et au caractère explicable des algorithmes afin de lutter contre l'inquiétude et l'incompréhension des citoyens vis-à-vis de l'intelligence artificielle.

Un des objectifs de ce projet est de développer une plateforme permettant la diffusion de ressources langagières afin de faciliter leur accès aux scientifiques. Il s'agit de proposer un nouveau modèle de partage des données et connaissances scientifiques qui favoriserait la communication entre les chercheurs, les fournisseurs de services et les citoyens.

2.2 L'équipe Sémagramme

Le Loria (Laboratoire lorrain de Recherche en Informatique et ses Applications), fondé en 1997 est une unité mixte de recherche commune au Centre national de la recherche scientifique (CNRS), à l'Université de Lorraine et à l'Institut national de recherche en sciences et technologies du numérique (Inria). Il a pour mission la recherche fondamentale et appliquée en sciences informatiques. Les travaux scientifiques y sont menés par 28 équipes structurées en 5 départements, ce qui représente un total d'environ 400 personnes. Les 5 départements sont les suivants :

1. Algorithmique, calcul, image et géométrie
2. Méthodes formelles
3. Réseaux, systèmes et services
4. Traitement automatique des langues et des connaissances
5. Systèmes complexes, intelligence artificielle et robotique

L'équipe Sémagramme appartient au département 4, c'est une équipe commune à l'Université de Lorraine, au CNRS et à l'Inria. Elle est dirigée par Philippe de Groote (directeur de recherche) et est composée d'un enseignant-chercheur, de deux chargés de recherche, d'un ingénieur et de cinq doctorants. Le but du projet Sémagramme est la définition et le développement de modèles, méthodes et outils, basés sur la logique, pour l'analyse sémantique d'énoncés et de discours en langue naturelle. Ceci inclut la modélisation logique de phénomènes pragmatiques liés à la dynamique du discours. Le projet est organisé autour de trois axes thématiques :

— Modélisation de l'interface Syntaxe-Sémantique
— Modélisation de la dynamique du discours.
— Développement de ressources linguistiques de base.

Autour de ces thématiques, la doctorants Chuyuan Li travaille sur des aspects Machine Learning, Samuel Buchel s'intéresse aux aspects psychologique dans le cadre du recueil de données pour la formation de corpus et Maria Boritchev travaille sur une vision compositionnelle des questions.

Mon travail est lié à la modélisation de la dynamique du discours puisqu'il consiste à travailler sur les relations discursives liant les actes de dialogue de conversations pathologiques.

À l'origine, SLAM était un projet de recherche de l'équipe Sémagramme. Il s'est ensuite développé et l'action exploratoire ODiM (Outils informatisés d'aide au Diagnostic des

Maladies mentales) de Inria est sa continuité. Elle est axée sur l'aide au diagnostic, un outil est développé dans ce cadre.

Enfin, le projet MePheSTO, en collaboration avec le *Deutsches Forschungszentrum für Künstliche Intelligenz* (DFKI), s'intéresse à l'identification des signes avant-coureurs à l'entrée dans la maladie. Ce projet est partenaire de SLAM - ODiM.

2.3 Les Archives Henri Poincaré

Les Archives Henri Poincaré - Philosophie et Recherches sur les Sciences et les Technologies, fondées en février 1992 par Gerhard Heinzmann, sont une unité mixte de recherche regroupant des membres de l'institut des sciences humaines et sociales (INSHS) du CNRS, de la composante Connaissances, Langage, Communication et Société de l'Université de Lorraine et de la Faculté des Sciences Historiques de l'Université de Strasbourg.

Elles sont composées d'une quarantaine de chercheurs et enseignants-chercheurs, de huit membres permanents formant l'équipe administrative et de deux ingénieurs en CDD contrat de projet. Le laboratoire accueille aussi de nombreux doctorants et des postdoctorants. Il est de plus très lié aux activités d'enseignement : plusieurs masters y sont adossés.

La personne qui a initié la création de ce laboratoire, Gerhard Heinzmann, avait travaillé pour sa thèse sur Poincaré qui avait une conception de la science liée à la philosophie. C'est avec cette même vision que travaille l'équipe des Archives : l'objectif est de mener des projets de recherche pluridisciplinaires, réunissant des scientifiques intéressés aussi à l'histoire et l'épistémologie de leurs disciplines mais aussi des philosophes, sociologues et historiens s'intéressant aussi à la science et pouvant guider les scientifiques dans leurs travaux de historiques et philosophiques.

3 Contexte du stage

La schizophrénie est une pathologie complexe et mal définie aux multiples manifestations cliniques : les symptômes caractéristiques permettant de la définir sont controversés et il est difficile de déterminer précisément les traits partagés par les individus présentant ce diagnostic.

L'objectif de ce projet est de mettre en évidence et de formaliser les dysfonctionnements conversationnels qui surviennent chez les personnes avec schizophrénie afin d'apporter des indices supplémentaires lors du diagnostic de cette pathologie. En étudiant des conversations pathologiques, on remarque que des difficultés apparaissent dans certains passages ; en modélisant ces extraits, on peut espérer comprendre ce qui pose problème comme expliqué dans (Amblard et al., 2014).

Comme les difficultés semblent liées à l'enchaînement des actes de dialogues, utiliser une modélisation qui d'une part explique le lien entre les actes et d'autre part possède des règles permettant de définir les dysfonctionnements conversationnels permet d'étudier précisément les disruptions. Un aperçu d'une représentation du discours correspondant à cela est donné dans (Busquets et al., 2001) : la *Théorie des Représentations Discursives*

Segmentées (S-DRT), théorie qui vise à expliquer l'interprétation d'un discours par la représentation que l'on s'en fait, y est présentée ainsi que les deux courants de recherche en analyse du discours dont elle s'inspire.

C'est sur la base de cette théorie que le modèle d'annotation du projet SLAM a été construit. Après la création d'un corpus pathologique, ce modèle est appliqué aux extraits présentant des discontinuités conversationnelles afin d'étudier les ruptures dans ces textes.

3.1 Modélisation de texte

Le projet SLAM vise à étudier l'enchaînement des actes de dialogue dans une conversation, la modélisation doit donc pouvoir rendre compte des actes en eux mêmes mais aussi de leur articulation les uns par rapport aux autres.

Cette analyse, dépassant le cadre de la phrase est au coeur de plusieurs courants de recherche. D'une part celui de **la sémantique dynamique** qui permet de représenter l'influence d'une phrase sur le contexte du discours et d'autre part **l'analyse du discours** qui cherche à mettre en évidence la structure du discours par son découpage en segments conversationnels et leur organisation relative. Ces deux courants ont donné naissance à la S-DRT (Lascarides & Asher, 2007), théorie qui essaye de modéliser le discours pour en expliquer son interprétation.

3.1.1 Deux approches du discours à la base de la S-DRT

Deux idées importantes pour l'interprétation du discours ont émergé dans les années 80. D'une part, la sémantique dynamique met en avant le rôle du contexte dans l'interprétation d'un segment conversationnel : un segment modifie le contexte global dans le quel il intervient, cette idée sera formalisée dans le cadre de la *Discourse Representation Theory*. D'autre part, un courant de recherche appelé Analyse du discours montre que l'analyse de la structure du discours est primordiale pour son interprétation.

Discourse Representation Theory

La *Discourse Representation Theory* (DRT) (Geurts et al., 2020) s'inscrit dans le courant de recherche de la sémantique dynamique qui est une extension au-delà de la phrase de la sémantique formelle.

La sémantique formelle cherche à proposer une représentation du sens de la phrase à travers la logique. Les Grammaires de Montague (Barwise & Moravcsik, 2014) en proposent une version opérationnalisée : le calcul du sens est basé sur les relations syntaxiques dans la phrase, ce qui correspond au principe de compositionnalité de Frege.

Mais ce modèle n'est pas suffisant notamment à cause de problèmes liés à la dynamique du discours : comme l'interprétation se restreint à la phrase, le reste du discours et donc des informations utiles au calcul du sens, est ignoré. Un exemple illustrant ce problème est celui des anaphores pronominales : si une phrase contient des pronoms, en analysant seulement son contenu et pas le contexte dans lequel cette phrase est amenée, il est impossible d'associer ces pronoms aux entités qu'ils représentent.

Une réponse à ce problème est proposée par (Heim, 2008) en proposant l'idée que la représentation d'une phrase n'est pas un prédicat logique mais une fonction qui modifie le contexte. Cette idée est modélisée par (Kamp & Reyle, 1993) dans le cadre de la DRT avec des boîtes, les *discourse representation structures* (DRS), représentant le contenu du discours. Ces boîtes sont mises à jour pour chaque nouveau segment conversationnel et les fonctions expliquant la fusion des boîtes sont les fonctions de changement de contexte.

Les DRS sont constituées d'une part d'un ensemble de référents discursifs, c'est à dire les entités évoquées dans la discussion, et d'autre part d'un ensemble de conditions représentant les informations contenues dans le discours.

La FIGURE 1 propose une représentation en DRT de la phrase « Un fermier possède un âne ». Elle montre qu'il y a deux référents discursifs x et y et trois conditions *fermier*, *possède* et *âne* où *fermier* concerne x, *âne* concerne y et *possède* concerne le couple (x, y). Dans cette modélisation, les conditions *fermier(x)* et *âne(y)* signifient qu'il y a deux entités x et y qui sont respectivement un fermier et un âne. La condition *possède(x, y)* signifie que l'entité x possède l'entité y, c'est-à-dire, avec les deux conditions précédentes que l'entité *fermier* possède l'entité *âne*.

Un fermier possède un âne.

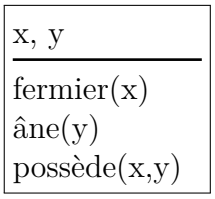


FIGURE 1 – Exemple de DRS

Au fur et à mesure du discours, la DRS globale se met à jour en combinant les DRS représentant chaque segment discursif. Plus explicitement, les référents discursifs qui réfèrent à un pronom sont substitués par l'individu correspondant : c'est la résolution des anaphores. Les boîtes sont la matérialisation des périmètres d'accessibilité des référents discursifs : on ne peut pas relier les référents discursifs correspondant à des pronoms à n'importe quel autre référent, des règles régissent l'accessibilité et les boîtes en sont la représentation.

La FIGURE 2 propose une représentation en DRT des phrases suivantes : Un fermier possède un âne. Il le bat.

En DRT on commence par modéliser la première phrase, ce qui nous ramène à la FIGURE 1. Puis on modélise la seconde phrase seule, ici les pronoms « Il » et « le » donnent deux référents discursifs a et b et le verbe *bat* donne la condition *bat(a,b)*. Contrairement à la modélisation de la première phrase, il n'y a pas de condition sur a ou sur b puisque la phrase ne nous apporte aucune information en elle-même, si ce n'est que les deux référents discursifs réfèrent à des entités masculines.

L'étape suivante est de fusionner ces deux modélisations pour former celle des deux phrases ensemble. On reporte les deux référents discursifs x et y ainsi que les trois conditions *fermier(x)*, *âne(y)* et *possède(x, y)* de la DRS de la première phrase puisqu'elle est complète. En revanche il faut désambigüiser la seconde phrase en rattachant a et b à des référents discursifs déjà explicités. C'est pourquoi on ajoute les conditions $a = x$ et

$b = y$, x et y sont bien accessibles puisqu'ils font partie des référents discursifs de cette DRS. Ainsi, a et b n'apparaissent pas dans les référents discursifs mais dans les conditions puisqu'ils sont en fait respectivement x et y. On conserve enfin *bat(a,b)* puisqu'il s'agit d'une information explicite contenue dans la seconde phrase.

Pour terminer la fusion, on supprime totalement a et b de la représentation en les remplaçant par x et y : *bat(a,b)* devient *bat(x,y)*. On supprime donc les conditions $a = x$ et $b = y$ qui deviennent inutiles.

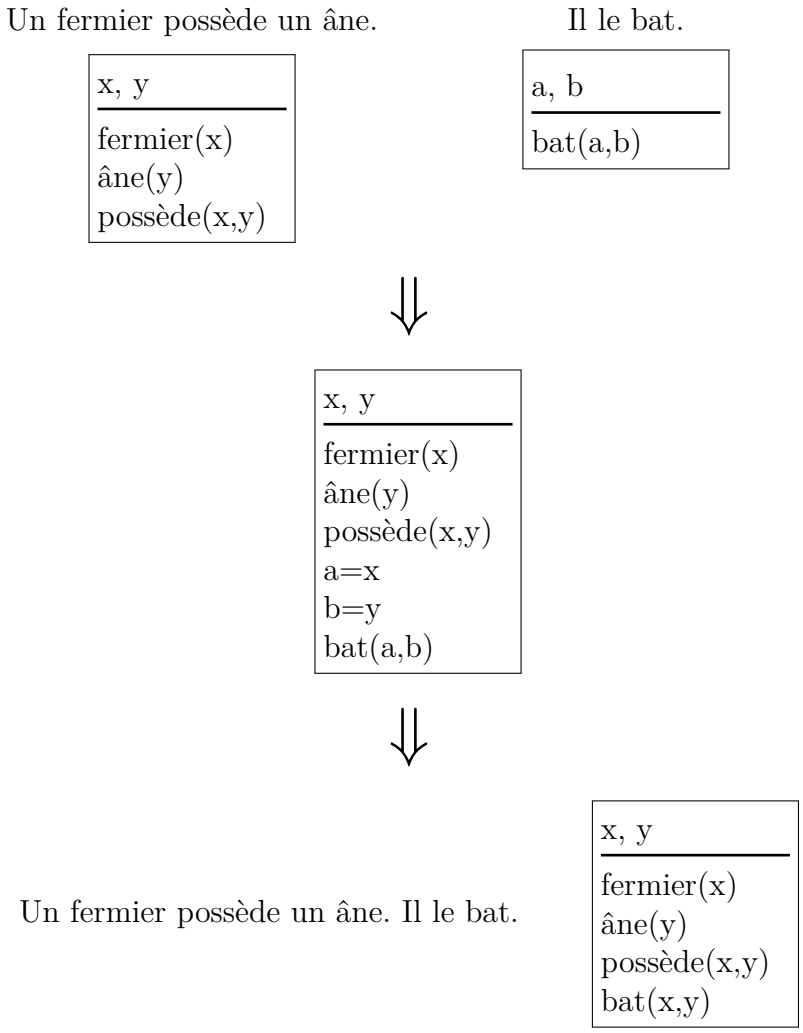


FIGURE 2 – Résolution d'anaphore pronominale en DRS

La DRT permet donc de prendre en compte le contexte et le dynamisme du discours dans sa modélisation. Elle ne permet en revanche pas d'analyser sa structure et les relations qui articulent ses parties.

Analyse du discours

Le second courant de recherche sur lequel se base la S-DRT, l'Analyse du discours, vise à décrire la macro-structure du discours et d'en analyser les relations entre ses différentes parties.

Cette théorie considère le discours comme formé de blocs homogènes de propositions élémentaires. L'homogénéité découle de l'enchaînement de segments linguistiques, ces segments ne se succèdent pas linéairement mais de manière structurée. Les hypothèses de départ sont les suivantes (Asher et al., 2003) :

— Tout discours cohérent a une structure.
— En tant que structure, il est possible de le formaliser.
— Une structure formelle permet de rendre compte de différents problèmes discursifs (résolution d'anaphores, structure temporelle, "l'emballage informationnel", structure thématique, entre autres).
— Il existe un ensemble de relations de cohérence qui rendent possible l'interprétation d'un discours.

La cohérence du discours peut s'observer grâce à l'existence de relations rhétoriques entre les segments conversationnels. Ces relations imposent une certaine structure au discours : certains segments discursifs sont directement accessibles aux suivants tandis que d'autres ne le sont plus et s'y référer sans le préciser explicitement provoque une rupture dans le discours.

L'ensemble de relations rhétoriques à utiliser pour analyser un texte ne fait pas consensus ce qui pousse à supposer qu'il existe différents types de discours et qu'à chaque type correspond une structure et donc des relations différentes.

L'Analyse du discours permet de mettre en évidence les relations entre les segments conversationnels d'un discours et d'analyser sa structure. Elle n'accorde en revanche pas d'importance, en général, au contenu propositionnel des segments, à l'inverse de la DRT. En couplant ces deux courants de recherche, une nouvelle théorie est apparue, permettant de rendre compte à la fois de la structure du discours par des relations rhétoriques entre les segments et du contenu de ces segments.

3.1.2 S-DRT et modélisation de discours

La *Segmented Discourse Representation Theory* (S-DRT) est une théorie d'interprétation du discours : elle permet de rendre compte de la structure du discours et de sa cohérence globale (Schlöder, 2019). Elle est basée sur le modèle de la DRT augmenté des théories d'Analyse du discours.

L'idée sur laquelle cette modélisation se base est la suivante : l'auditeur se construit une représentation mentale du discours qui évolue après chaque phrase de sorte que tous les segments sont reliés entre eux au sein d'une structure qui doit être logique pour que le discours soit compris.

La S-DRT suit le même principe de construction que la DRT : chaque nouveau segment est représenté sous forme de DRS puis fusionné à la structure préexistante notamment en faisant correspondre les référents discursifs qui représentent la même chose. Enfin, comme en Analyse du discours, on définit une relation entre les segments qui explique le rôle au sein du discours de chacun d'entre eux.

Fonctionnement de la S-DRT

La S-DRT (Asher et al., 2003) permet de modéliser un discours sous forme de *Segmented Discourse Representation Structures* (SDRS) qui ressemblent aux DRS mais où les référents discursifs sont remplacés par des étiquettes référant aux actes de langage. Les segments vérifient une condition sous forme de DRS et sont aussi reliés entre eux par des relations rhétoriques permettant de mettre en évidence l'enchaînement des idées mais aussi le rôle de chaque segment dans le discours. Les relations rhétoriques sont de deux types, coordonnantes et subordonnantes. Dans les relations coordonnantes, les deux segments sont au même niveau, aucun ne domine l'autre : dans une représentation sous forme de graphe des segments du discours ils sont reliés par une ligne horizontale. Dans les relations subordonnantes, un segment domine l'autre : dans une représentation sous forme de graphe des segments du discours ils sont reliés par une ligne verticale.

Pour créer cette représentation, chaque proposition (ou segment) du discours est formulée en DRT avant d'être fusionnée à la représentation globale.

Les points d'attachement possible sont ensuite déterminés dans la SDRS déjà construite. Ceci se fait selon certaines règles et notamment celle de la Frontière droite : les seuls noeuds accessibles sont le dernier noeud ajouté au graphe et ceux qui dominent celui-ci.

Enfin on met à jour la SDRS globale en insérant le constituant que l'on vient de construire, en résolvant les sous-spécifications puis en ajoutant des relations rhétoriques pour lier les items entre eux.

La FIGURE 3 propose une représentation en S-DRT des phrases suivantes : Un fermier possède un âne. Il le bat.

Les deux phrases sont d'abord représentées sous forme de DRS, on obtient donc les mêmes DRS qu'à la première étape de la FIGURE 2 : K_{π_1} (respectivement K_{π_2}) est la condition vérifiée par π_1 (respectivement π_2).

Comme on cherche à représenter le contenu des deux phrases et le lien entre elles, on crée une boîte dont les segments sont π_1 et π_2, où π_1 vérifie K_{π_1} et π_2 vérifie K_{π_2}, ce sont les deux conditions $\pi_1 : K_{\pi_1}$ et $\pi_2 : K_{\pi_2}$. Le segment π_2 apporte des informations qui complètent ce qui est dit dans π_1, c'est donc une élaboration, d'où la condition Elaboration(π_1, π_2).

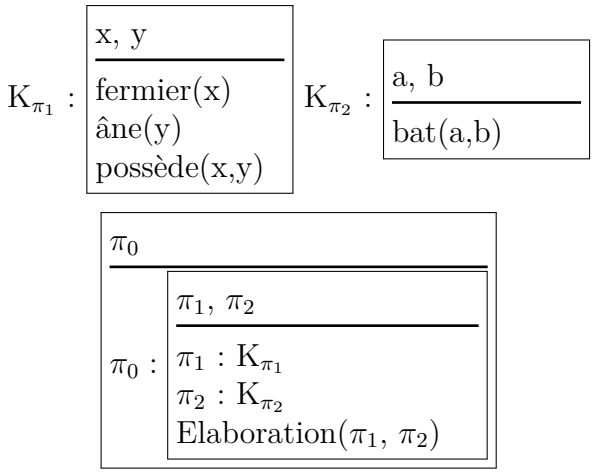
Pour terminer cette représentation, on insère cette boîte dans une autre étiquetée π_0 où elle devient la condition vérifiée par π_0. π_0 représente les deux phrases de manière globale, ce qui permettrait par la suite de rattacher un nouveau segment à π_2 ou π_0 selon le contenu du nouveau segment.

Une représentation sous forme de graphe correspond à ces boîtes, les segments π_1 et π_2 sont reliés verticalement par une Élaboration, il s'agit en effet d'une relation subordonnante ou π_1 domine π_2. Cette représentation est plus simple à lire, notamment quand il s'agit de la modélisation d'un texte plus long.

La mise a jour de la structure globale lors de l'ajout d'un acte de dialogue est une phase complexe de cette modélisation, c'est elle qui rend compte de l'aspect dynamique du discours. Plusieurs problèmes s'y posent et deux d'entre eux sont décrits ici de manière exhaustive.

Deux des problèmes à résoudre lors de la mise à jour

π_1. Un fermier possède un âne.
π_2. Il le bat.



(a) Représentation en SDRS

(b) Graphe correspondant

FIGURE 3 – Exemple de SDRS

π_1. Jean était au restaurant.
π_2. Il a commandé une salade.
π_3. Puis il a mangé une assiette de pâtes.
π_4. Mais il ne l'a pas trouvée à son goût.

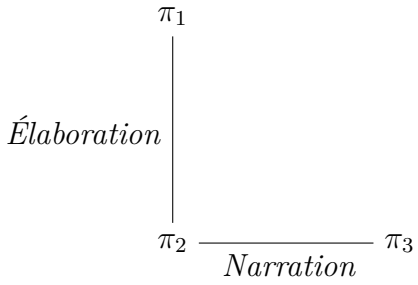


FIGURE 4 – Résolution d'anaphore pronominale

La DRT permet de résoudre à elle seule certains problèmes de la mise à jour en prenant en compte le contexte dans lequel celle-ci s'inscrit. Mais il arrive que le contexte soit trop large pour permettre la résolution de toutes les sous-spécifications. Dans ces cas, l'ajout de relations rhétoriques entre les segments conversationnels permet de réduire le périmètre.

Anaphores pronominales

Quand plusieurs référents discursifs peuvent correspondre à un pronom, ce sont les relations qui permettent de définir lesquels sont accessibles. Par exemple sur la FIGURE 4, dans π_4, « l' » peut référer à la salade ou à l'assiette de pâtes d'après le genre et le nombre de « trouvée ». Mais la construction du discours fait que π_2 et donc « salade » n'est plus accessible : c'est la règle de la frontière droite. Les seuls segments accessibles sont le dernier segment fusionné et ses parents dans le graphe soit ici π_3.

Anaphores temporelles

Dans un discours, les événements décrits ne le sont pas toujours dans l'ordre chronologique. Les relations liant les segments doivent prendre en compte ce fait et il faut alors

se baser sur ses propres connaissances en plus des règles de la S-DRT pour les choisir. Ainsi dans (1), les événements semblent décris dans le bon ordre mais pas dans (2) car nos connaissances nous indiquent que pousser cause la chute, il nous est donc assez simple de choisir *Cause* ou *Conséquence* comme relation entre les deux phrases.

(1) Jean est tombé. Max l'a aidé.
(2) Jean est tombé. Max l'a poussé.

La S-DRT propose donc une modélisation permettant de rendre compte de l'organisation et du rôle des actes de dialogue les uns par rapport aux autres tout en fournissant des règles pour leur rattachement. Cette théorie est donc une bonne base pour modéliser les conversations pathologiques. Pour l'adapter aux besoins du projet, un ensemble de relations rhétoriques a été défini, notamment au cours de projets tutorés (Huber & Laurier, 2017) et (Biver et al., 2018). Ces relations, décrites dans la Table 1, sont assez précises pour les phrases déclaratives mais peu pour les paires question-réponses, or préciser le rôle des questions et des réponses dans le dialogue peut aider à la compréhension globale du dialogue.

3.1.3 Précision des relations pour les paires questions-réponses

Les paires question-réponse sont un phénomène complexe à analyser et pourtant très présent dans les dialogues. Des relations rhétoriques précises pour lier les questions et réponses entre elles ainsi qu'au reste du dialogue peuvent donc permettre une meilleure appréhension de leur contribution à la conversation.

C'est dans cette optique que le schéma d'annotations proposé dans (Cruz Blandón et al., 2018) a été développé. Ce schéma permet d'indiquer l'étendue de la question ou réponse mais aussi de définir précisément son rôle dans la discussion. L'objectif étant de le rendre utilisable tant par des annotateurs humains que des algorithmes de Machine Learning, les critères permettant de repérer et classifier les questions et réponses ont été choisis de sorte qu'ils puissent être appliqués conformément par les uns et les autres. Par exemple, les questions sont repérées grâce au point d'interrogation, or il arrive que des segments conversationnels aient un sens interrogatif sans présenter de point d'interrogation, ne serait-ce qu'à cause d'un oubli dans la transcription. Dans ces cas, le contexte et la prosodie permettent d'identifier une question ; c'est une analyse qui aurait trop complexe pour la visée de ce schéma d'annotation mais qui peut être envisagée lors de l'annotation en S-DRT.

Ce schéma apporte donc une bonne base pour préciser les relations rhétoriques de type questions-réponses.

3.2 Méthode de travail pour le projet

Le projet SLAM s'intéresse à l'analyse de conversations pathologiques, il nécessite donc la création et l'étude de corpus dont les données sont recueillies dans des centres hospitaliers. Ces corpus sont constitués de tests neuropsychologiques, de différents enregistrements par des dispositifs spécifiques comme des *eye-trackers* ou des EEG et de transcriptions d'enregistrements audio d'entretiens avec un psychologue. Les transcrip-

tions sont ensuite analysées grâce à plusieurs types d'annotations, automatiques et manuelles. Les extraits présentant des ruptures sont de plus modélisés sous forme d'un graphe où les segments d'interventions sont reliés par des relations rhétoriques.

3.2.1 Recueil des données

La constitution de corpus (Amblard et al., 2019b) est primordiale pour le projet. C'est un processus assez complexe d'une part car l'organisation d'entretiens avec des patients schizophrènes est très encadrée mais aussi à cause des difficultés posées par la constitution d'un corpus témoin.

Entretiens et transcriptions

Les entretiens avec le psychologue varient d'un patient à l'autre. Le rôle du psychologue est de maintenir l'échange, ce qui chez certains patients l'oblige à faire une grande partie de la discussion alors que chez d'autres l'échange est plus équilibré.

Ces entretiens sont enregistrés puis transcris manuellement afin d'obtenir une version lisible des entretiens. Le but étant de transcrire exactement ce qui est entendu, seule la conversation est conservée : les mouvements et déplacements ne sont pas transcris et seuls les bruits nécessaires à la compréhension retenus. Un guide d'annotations a été rédigé pour aider les transcripteurs dans leur tâche.

Difficultés liées au recueil des données

Comparer le corpus témoin au corpus pathologique n'est pas une tâche simple. En effet, les entretiens avec les patients ont lieu dans le cadre médical, les traitements médicamenteux auxquels ils sont soumis varient d'un patient à l'autre. Ces traitements peuvent induire des effets secondaires difficiles à anticiper et qui constituent une claire dissymétrie par rapport au groupe témoin. De plus, les patients ont souvent un QI moins élevé et/ou un niveau d'étude plus bas que la population générale dans la même tranche d'âge.

3.2.2 Annotation des transcriptions

A la lecture du corpus, il semble qu'il y ait des dysfonctionnements à un niveau abstrait : au niveau sémantico-pragmatique et dans la planification de l'interaction. Mais ce ne sont pas nécessairement les seuls niveaux de dysfonctionnements, il faut donc vérifier si aux autres niveaux, plus concrets, tout fonctionne correctement. Des outils ont été développés pour analyser ces niveaux et il est donc intéressant de se tourner vers des annotations automatiques pour les analyser.

Annotations automatiques

Cette première étape a pour but d'analyser la qualité de la production langagière afin de s'assurer qu'elle est semblable dans les deux groupes.

Une analyse syntaxique est d'abord effectuée. Aucun analyseur ne fait consensus pour le français parlé notamment car la notion de phrase à l'oral n'est pas triviale. C'est UDPipe, un logiciel libre, qui a été utilisé.

L'identification des disfluences, c'est à dire de réalisations orales qui rompent la continuité syntaxique, permet d'effectuer une comparaison supplémentaire. L'outil choisi pour cette analyse est Distagger, un logiciel libre qui permet de repérer les disfluences dans des transcriptions orales et notamment les interjections d'hésitation, la reprise à l'identique d'un mot ou d'un groupe de mots, l'autocorrection immédiate et l'interruption de morphème en cours d'énonciation.

Enfin, une analyse morpho-syntaxique a été menée grâce à l'outil libre MElt dans une version entrainée sur des ressources suffisamment proches du corpus pour lui correspondre.

S'il y a légèrement plus de disfluences dans le groupe des patients schizophrènes, aucun groupe ne se distingue en termes de diversité et de richesse syntaxique.

Annotations manuelles

Le niveau sémantico-pragmatique est ensuite étudié grâce à des annotations manuelles. Comme la production de ces annotations est coûteuse, il n'est pas possible de les produire pour l'intégralité des transcriptions. La première étape est donc de repérer dans chaque conversation les extraits présentant des ruptures décisives (Amblard et al., 2014). Puis la production d'une représentation formelle de ces extraits est discutée.

Ces deux étapes sont réalisées par des experts car c'est une analyse complexe qui nécessite la prise en compte de la conversation dans son ensemble et notamment des propriétés de planification du dialogue.

Pour rendre compte de cette analyse, les entretiens sont ensuite modélisés.

Modèle retenu

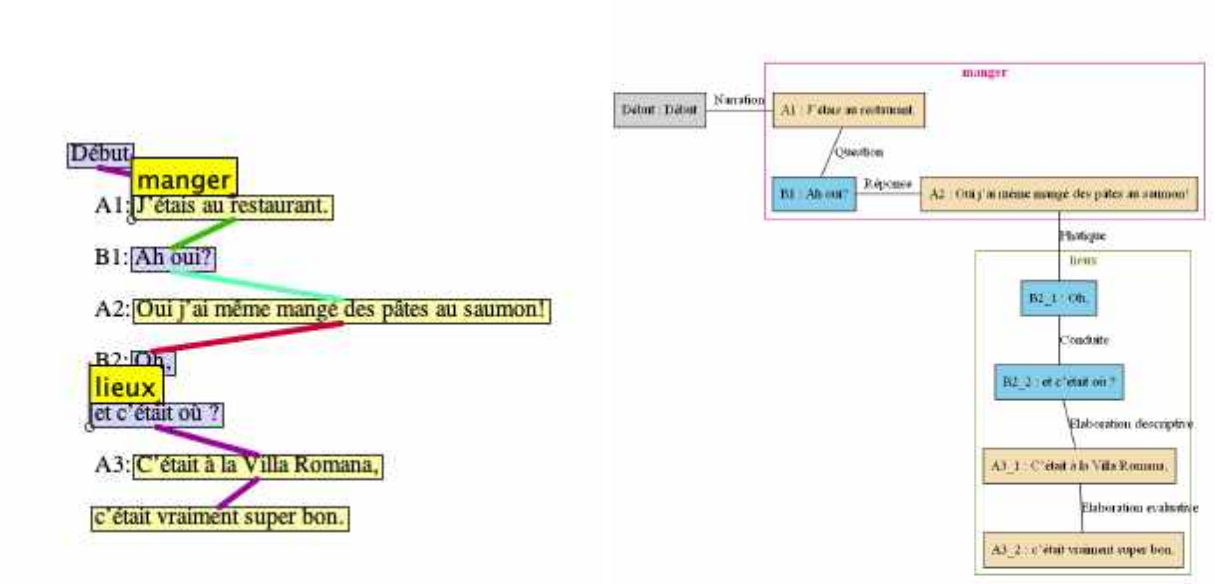
La formalisation retenue pour le projet SLAM s'inspire de la S-DRT. Il s'agit d'une part d'un formalisme plus simple permettant de lire le dialogue tout en analysant sa structure. D'autre part, la S-DRT est une théorie représentationnelle du discours et non du dialogue, les relations rhétoriques retenues pour annoter le corpus sont donc légèrement différentes de celles envisagées dans le cadre de la S-DRT.

L'ensemble des relations rhétoriques a évolué au fur et à mesure du projet et notamment dans le cadre de projets tutorés liés aux annotations. Le projet tutoré (Huber & Laurier, 2017) avait pour but de mener une campagne d'annotations. Dans ce cadre, une réflexion a été menée sur les différentes relations possibles à proposer aux annotateurs et les annotations de la Table 1 ont été retenues. Les relations ont été regroupées par catégories.

Annoter une transcription consiste dans un premier temps à relier les unités, qui sont des phrases ou des segments de phrases, grâce aux relations rhétoriques de la Table 1. Chaque unité doit être reliée à une unité qui la précède mais pas nécessairement celle qui la précède directement. Dans un second temps, il faut préciser les thèmes associés à chaque partie.

Narrations	Narration
Élaborations	Élaboration descriptive Élaboration prescriptive Élaboration évaluative Contre-élaboration
Méta	Conduite Phatique Méta-Question
Questions / Réponses	Question Réponse

TABLE 1 – Premier ensemble de relations discursives



(a) Annotation sur Glozz (b) Annotation sous forme de graphe

FIGURE 5 – Exemple d'annotation d'une transcription sur Glozz

La FIGURE 5 est un exemple d'annotation. Sur la figure de gauche, on peut voir que les interventions de chaque locuteur sont surlignées de couleurs différentes. Les relations sont les traits qui relient deux unités, celles de même couleur appartiennent à la même catégorie. Les thèmes sont modélisés par les drapeaux jaunes, un thème s'étend de l'unité portant le drapeau à celle précédent le drapeau suivant comme on peut le remarquer sur la modélisation sous forme de graphe sur l'image de droite. Sur cet exemple, chaque unité est reliée à celle qui la précède directement mais ce n'est pas toujours le cas.

L'année suivante, une nouvelle campagne d'annotation basée sur les mêmes relations a été réalisée dans le cadre du projet tutoré (Carletti et al., 2019). Les données recueillies ont ensuite été analysées et au vu des remarques des différents annotateurs, l'ajout d'une relation neutre signifiant que deux unités sont en relations mais que l'annotateur n'en connaît pas le type ou qu'il n'y a pas de relation appropriée a été suggérée. La liste de relations disponibles pourrait donc être complétée pour que les annotations gagnent en

précision.

Résultats actuels

Le premier corpus étudié (Amblard et al., 2011) comprend **30 entretiens** avec **14 patients schizophrènes paranoïdes**, **8 patients schizophrènes désorganisés** et **8 sujets dans un groupe contrôle**.

Postulats d'étude

Pour comprendre ce qu'il se passe dans la conversation, deux postulats ont été faits (Rebuschi et al., 2013) :

— Le patient et le psychologue ont une représentation mentale différente de l'interaction
— Le contenu sémantique des paroles du patient est consistant : sa représentation du monde, telle qu'elle se construit au travers au fil de la conversation, est logique.

Ruptures identifiables

Deux types de ruptures ont été identifiés dans ce corpus. D'une part des ruptures de la frontière droite : un segment conversationnel ne peut être rattaché qu'au dernier nœud créé ou aux parents de celui-ci, sinon il y a rupture de la structure conversationnelle. Et d'autre part des remontées dans la structure : si un segment est rattaché très haut par rapport au dernier nœud créé sans que la conversation présente de marqueurs de type « Par ailleurs » ou « Pour revenir à ... » alors il y a aussi rupture car un locuteur change de sujet sans clore le précédent proprement.

De plus, on remarque que les ambiguïtés sont des catalyseurs de dysfonctionnements : le thème choisi pour désambigüiser une expression change brusquement, sans clôture propre du thème précédent.

Après étude du corpus, 403 séquences d'interactions ont été obtenues :

— Plus de 30% présentant des discontinuités chez les patients schizophrènes
— Moins de 2% chez le groupe contrôle

Si les résultats semblent mettre en avant une différence significative entre les conversations pathologiques et celles du groupe contrôle, il faut prendre en compte le fait que la cohérence d'un dialogue et la présence de ruptures sont des faits subjectifs. La seule possibilité pour justifier ces résultats est d'atteindre un accord intersubjectif entre les annotateurs.

3.2.3 Vérification des résultats

Des campagnes d'annotations ont été mises en place dans le cadre de projets tutorés pour vérifier les résultats. Des annotateurs naïfs (non experts et ne connaissant pas l'objectif des annotations) ont annotés plusieurs extraits de dialogues pathologiques ou non (textes témoins). Le fait qu'ils ne soient pas au courant de l'objectif des annotations permet de limiter les biais.

Un guide d'annotations était à leur disposition expliquant précisément chaque relation et les illustrant avec des exemples. Ces annotations ont ensuite été comparées entre elles ainsi qu'avec les annotations des experts.

La FIGURE 6 est une représentation sous forme de graphe d'un extrait de conversation pathologique nommé Provocation. Il s'agit de la fusion de toutes les annotations pour ce texte : seules les annotations majoritaires (le seuil est de 15%) sont visibles. Sous chaque relation, le nombre d'annotateurs l'ayant choisie est noté. De plus, au niveau des unités, le nombre d'annotateurs ayant signalé un changement de thème est noté en rouge. Cet extrait a été annoté par 21 personnes. Les annotations les plus consensuelles ici son Narration, Élaboration descriptive, Question, Réponse et Phatique.

Plus globalement, en étudiant tous les extraits de cette campagne, ce sont les relations Question, Réponse et Phatique qui font le plus consensus. Les annotateurs sont en général d'accord sur les emplacements des relations : les endroits qui posent problèmes correspondent à des unités dont il est difficile de comprendre le sens.

Quand à la comparaison entre les annotateurs naïfs et les experts, les premiers ne trouvent pas plus de ruptures de la frontière droite dans les extraits de conversations pathologiques que dans les textes témoins, contrairement aux seconds. En revanche, ils trouvent plus de remontées dans la structure pour certains des extraits de conversations pathologiques.

Les analyses menées sur ce premier corpus mettent en évidence la présence de ruptures de la frontière droite plus nombreuses dans les conversations pathologiques. Un second corpus est en cours de transcription et c'est sur deux des conversations qui le composent que j'ai travaillé par la suite.

4 Travail effectué pendant le stage

Mon travail a consisté tout d'abord à prendre en main les logiciels et les scripts Python rédigés dans (Biver et al., 2018) afin d'être capable ensuite de les utiliser pour annoter deux nouvelles transcriptions.

Les transcriptions décrites ci-après sont celles d'entretiens réalisés en 2019 à Tizi Ouzou.

L'entretien HA-R est constitué d'environ 700 interventions. Il se divise en trois partie. Pendant un tiers de la conversation, le psychologue explique l'objectif de l'étude au patient, celui-ci est très intéressé par le sujet : il parle de ses propres connaissances et pose des questions sur les mots qu'il ne comprend pas. Puis le psychologue lui demande comment il est arrivé à l'hôpital et le sujet change : plusieurs thèmes s'enchaînent et la conversation semble moins claire que précédemment. Durant cet échange, le climatiseur fait du bruit, ce qui fait réagir les deux interlocuteurs qui mettent de côté leur discussion pour parler de cela, ils reviennent finalement à leur conversation après avoir clos le sujet. Puis après quelques minutes de discussion le patient demande au psychologue de lui parler de la « nouvelle science » et la conversation se poursuit comme au début de l'échange.

La partie qui semble moins claire représente un quart de la discussion complète, c'est

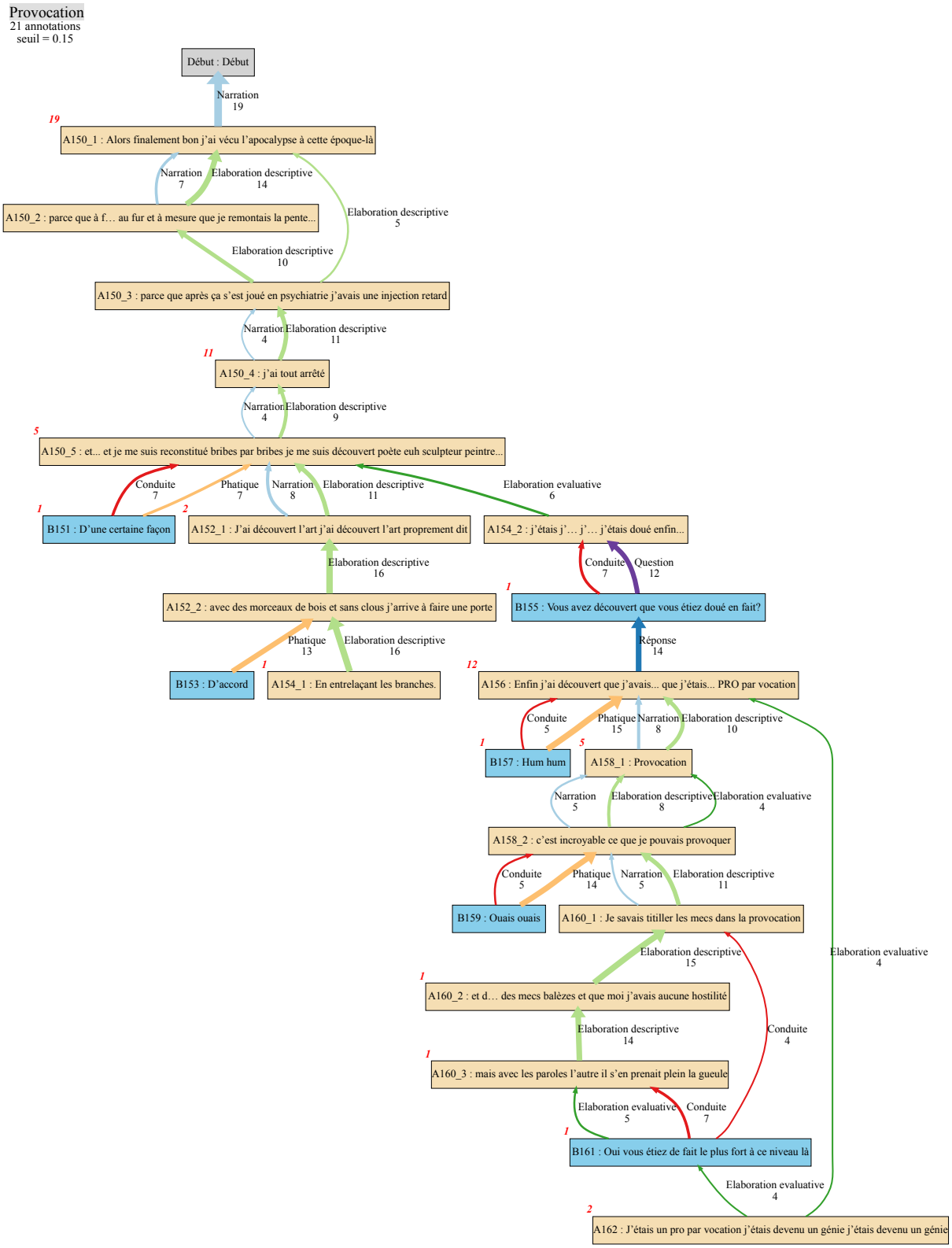


FIGURE 6 – Graphe total des annotations de l'extrait Provocation

celle que j'ai annotée plus précisément, je l'ai donc découpée en segments conversationnels ce qui m'a donné 237 unités, soit 236 annotations.

L'entretien Am-A est constitué d'environ 600 interventions qui une fois découpées en segments conversationnels donnent 694 unités à annoter, soit 693 relations à créer. Le premier tiers de la conversation est centrée sur le protocole : le psychologue explique pourquoi il est là et cela donne l'occasion au patient de poser des questions ou au psychologue de l'interroger sur ses troubles. Il y a donc des digressions mais la discussion revient au protocole à quatre reprises. Puis la conversation est orientée sur le patient, ses études ou la vie qu'il mène hors de l'hôpital par exemple. Un problème avec la porte survient ensuite et les deux interlocuteurs s'interrogent à ce sujet avant que le psychologue ne rebondisse sur une intervention du patient pour réorienter la discussion et continuer de parler de sa vie notamment hors de l'hôpital. Enfin le patient interrompt la discussion en expliquant qu'il souhaite partir car il est fatigué, la discussion tourne autour des raisons de cette fatigue pendant une centaine d'unités avant de se terminer.

Comme j'ai reçu la transcription Ha-R en premier, j'ai travaillé dessus jusqu'à ce que mon modèle d'annotation soit satisfaisant pour un échange long. Je ne l'ai annotée qu'en partie car c'est ce qui avait été fait lors de l'étude du précédent corpus. Cela dit, l'extrait a été choisi car il semblait légèrement différent du reste de la conversation et non pas car des ruptures manifestes y apparaissaient, contrairement aux extraits choisis dans le précédent corpus, ce qui fait qu'il est plus long.

Or l'étude d'un extrait aussi long a mis en évidence des relations entre les différents thèmes de la conversation, ce qui était difficilement repérable dans les précédents extraits. En plus de la modélisation du dialogue segment par segment, une modélisation des boîtes thématique s'inspirant aussi de la S-DRT a été mise au point. Ce second modèle est basé sur les thèmes, or ceux-ci ne peuvent être mis en évidence que sur de longues conversations, c'est pourquoi, une fois les modèles mis au point j'ai annoté l'intégralité de la transcription Am-A.

4.1 Prise en main des logiciels et des scripts python

Deux logiciels sont utilisés pour analyser les entretiens. Tout d'abord ELAN, pour les transcrire et, dans mon cas effectuer un premier travail d'annotations. Puis Glozz, pour annoter précisément les actes de dialogue. De plus, des étudiants ont produit lors d'un projet tutoré des scripts Python permettant de construire les graphes correspondant aux annotations de type S-DRT de Glozz et d'étudier le degré d'accord entre différents annotateurs.

4.1.1 Annotations de texte avec ELAN

ELAN est un logiciel permettant la création d'annotations complexes sur les ressources vidéo et audio. Il est possible d'ajouter un nombre illimité d'annotations à la ressource ainsi que de les disposer sur différentes couches appelées *tiers*. Les annotations sont des chaînes de caractères, elles peuvent donc représenter n'importe quoi : un mot, une phrase, un commentaire ou même la transcription textuelle de la ressource.

ELAN est de plus capable de générer des statistiques sur les annotations réalisées sur un fichier et même de comparer différent fichiers.

Dans le cadre du projet, les entretiens sont enregistrés puis transcris à l'aide d'ELAN comme sur la FIGURE 7 : un *tiers* est consacré à chaque participant de la discussion. Les annotations sont ensuite exportées afin de créer un fichier texte : c'est la transcription de l'échange.

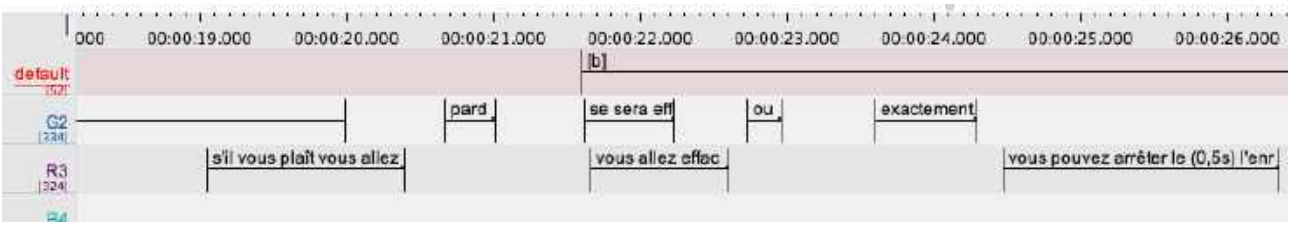


FIGURE 7 – Une transcription sur ELAN

Dans ces transcriptions figurent des indications telles que les temps de silence et les différents bruits entendus dans l'enregistrement identifiés par des lettres entre crochets comme on peut le voir sur la FIGURE 8.

002 default [b]

003 R3 s'il vous plaît vous allez effacer

004 G2 pardon

005 default [b]

006 G2 se sera effacé

007 R3 vous allez effacer

008 G2 ouais

009 G2 exactement ouais

010 R3 vous pouvez arrêter le (0,5s) l'enregistrement

011 G2 [i] ah on euh (0,5s) pas arrêter l'enregistrement mais on l'efface (0,5s) on l'efface

FIGURE 8 – Extrait d'une transcription

4.1.2 Annotations de texte avec Glozz

Glozz est un logiciel libre développé par plusieurs équipes de recherche permettant d'annoter des textes et de visualiser les relations créées.

Trois types d'annotations existent comme on le voit sur la FIGURE 9 : les unités, les relations et les schémas. On peut choisir leur couleur dans les onglets style.

Les unités permettent de surligner des morceaux de texte comme sur la FIGURE 10, on peut en choisir la couleur et le nom et éventuellement ajouter des informations spécifiques grâce à l'onglet Features. Les relations permettent de relier deux unités comme sur la FIGURE 11, on peut là encore choisir la couleur et le nom de chaque relation et ajouter des informations. Enfin les schémas permettent de relier plusieurs unités et/ou relations. Ce dernier type d'annotations ne sera pas utilisé ici.

Glozz utilise des fichiers écrits dans des langages tels que le XML et le CSS, il est donc possible de créer ses propres schémas d'annotation. Un manuel fournit toutes les explications pour cela et donne aussi des indications pour automatiser la création de corpus.

Automatisation de la préparation des fichiers Glozz

Pour annoter un texte dans Glozz avec le modèle retenu pour ce projet, il faut le découper en segments conversationnels puis les numéroter pour simplifier l'analyse. Après avoir importé le texte découpé dans Glozz, il faut créer les unités en surlignant les segments un par un avec la couleur correspondant à l'interlocuteur.

Ce travail de préparation était réalisé à la main lors de l'étude du précédent corpus mais les extrait annotés étaient bien plus courts. C'est pourquoi j'ai choisi d'automatiser la numérotation des unités dans le fichier texte et leur création dans Glozz grâce à des scripts Python.

Numérotation des segments conversationnels

Avant de pouvoir numéroter les lignes, il faut segmenter la discussion à la main. Pour signifier les différents segments, il suffit de sauter une ligne entre deux segments conversationnels d'une même intervention et deux lorsque l'interlocuteur change. Un seul saut de ligne dans les deux cas aurait été possible aussi mais faire la distinction entre les interlocuteurs manuellement simplifie le code et dans la mesure où la segmentation se fait de toute façon à la main, j'ai pensé que ce choix ne posait pas de problème.

Il y a dans les transcriptions des lignes et des caractères inutiles aux annotations comme les lignes default ou des suites de (xxx). J'ai choisi de les supprimer à la main en les cherchant grâce à des expressions rationnelles. Ce processus pourrait aussi être automatisé si une liste des motifs inutiles était dressée. Le faire directement avec les fonctions d'un éditeur de texte ne m'a pas pris beaucoup de temps, c'est pourquoi je n'ai pas automatisé ce processus.

Pour numéroter les segments, le script lit le texte ligne à ligne et assigne les bons numéros grâce à des compteurs. En lui indiquant en entrée les deux interlocuteurs dans l'ordre de première apparition, il est facile de déterminer qui parle en prenant en compte les lignes vides. En revanche, ce choix de codage fait que la présence d'un autre interlocuteur rend le code inutilisable.

Création des unités Glozz

Le fichier lu par Glozz pour retrouver les annotations et les unités pour un texte et

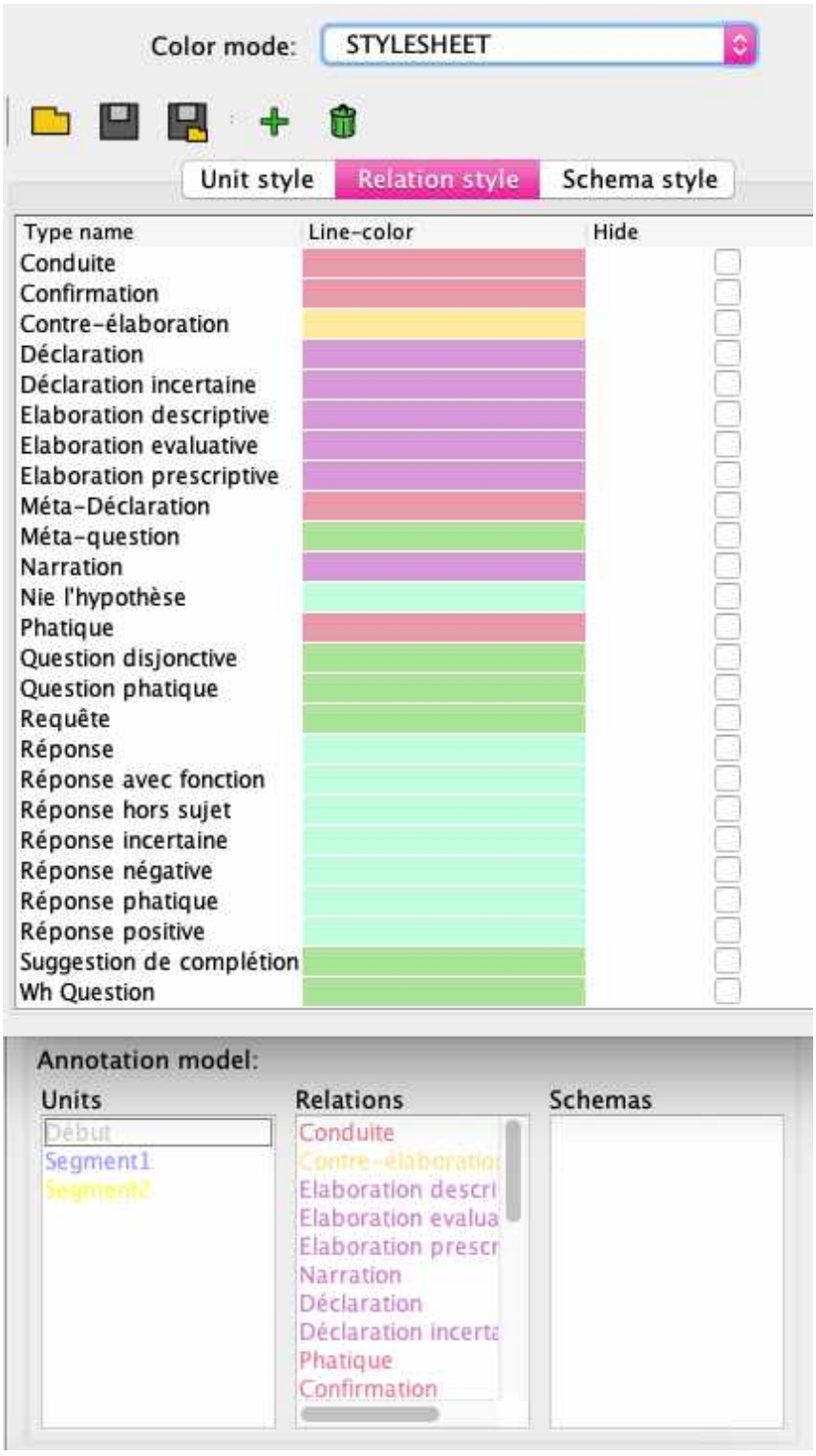


FIGURE 9 – Modèle d'annotations

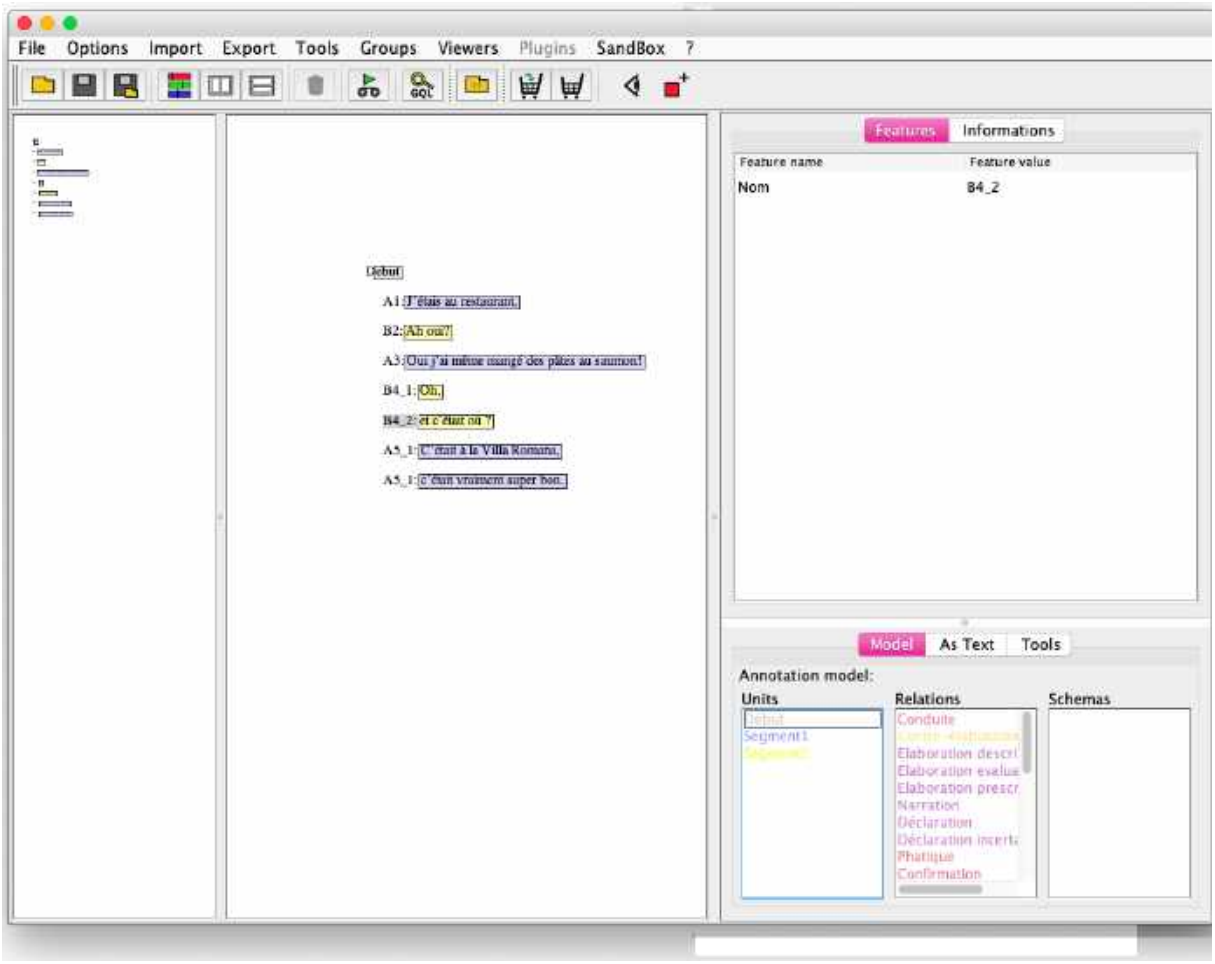


FIGURE 10 – Annotation des unités

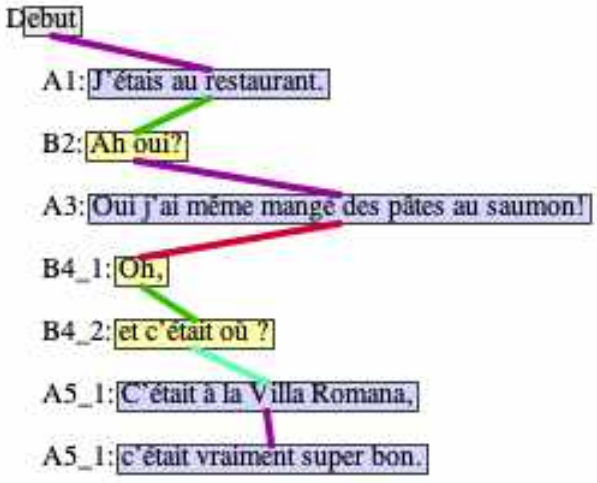


FIGURE 11 – Création des relations

un fichier .aa. C'est en fait un fichier écrit en xml, il est donc assez simple de le créer automatiquement.

Nous utilisons trois balises types d'annotations : les unités (<unit>), les relations (<relation>) et les drapeaux (<flag>) indiquant les thèmes. Les relations et les drapeaux sont ajoutés manuellement lors de l'annotation mais la création des unités peut être automatisée puisqu'il s'agit simplement de créer une unité par segment conversationnel avec une couleur par interlocuteur.

Listing 1 – Arborescence du fichier .aa lu par Glozz

```
0: annotations [-]
    1: metadata [corpusHashcode]
    1: unit [id]
        2: metadata [-]
            3: author [-]
            3: creation-date [-]
            3: lastModifier [-]
            3: lastModificationDate [-]
        2: characterisation [-]
            3: type [-]
            3: featureSet [-]
        2: positioning [-]
            3: start [-]
                4: singlePosition [index]
            3: end [-]
                4: singlePosition [index]
                4: feature [name]
    1: relation [id]
        2: metadata [-]
            3: author [-]
            3: creation-date [-]
            3: lastModifier [-]
            3: lastModificationDate [-]
        2: characterisation [-]
            3: type [-]
            3: featureSet [-]
                4: feature [name]
        2: positioning [-]
            3: term [id]
    1: flag [id]
        2: metadata [-]
            3: author [-]
            3: creation-date [-]
            3: lastModifier [-]
            3: lastModificationDate [-]
        2: characterisation [-]
            3: comment [-]
        2: positioning [-]
            3: singlePosition [index]
```

Pour créer une unité, il faut commencer par créer son identifiant. Glozz utilise la date de création de l'unité sous forme de timestamp comme identifiant et, même si n'importe quel type d'identifiant pourrait convenir, j'ai choisi de faire de même dans la mesure où c'est une solution simple et efficace. En revanche, compte tenu de la vitesse d'exécution d'un script Python par rapport à celle de la création manuelle de chaque unité, les 13 chiffres des timestamp Glozz n'étaient pas suffisant : plusieurs unités se retrouvaient avec le même identifiant. Après plusieurs essais, j'ai jugé qu'avec 16 chiffres la différence entre les identifiants de deux unités consécutives était suffisante pour éviter ce problème.

Le texte correspondant à chaque unité est identifié par sa position (début et fin) dans un fichier .ac qui contient le texte sans sauts de ligne. J'ai donc créé les unités les unes

après les autres en lisant le fichier texte numéroté ligne à ligne : un compteur permettait de déterminer la position de début et la longueur de la ligne donnait la position de fin, le compteur se mettant à jour après la création d'une unité.

Il faut créer deux types d'unités : les premières permettent la découpe du texte par Glozz et les autres correspondent au surlignage des segments conversationnels. La distinction des deux types se fait grâce aux balises <type>. Les unités de découpage du texte sont de type paragraph et les autres de type segment1 ou segment2 selon l'interlocuteur. De plus l'auteur des unités de découpage du texte est TXT_IMPORTER alors que celui des segments est choisi par l'annotateur comme le ferait Glozz si tout était fait à la main.

Enfin j'ai choisi pour les segments de ne surligner que la partie correspondant au texte de chaque unité : l'identifiant (G12_2 par exemple) n'est pas surligné. En revanche, dans l'élément <featureSet> qui permet d'ajouter des précisions à une annotation, l'élément <feature> Nom contient cet identifiant comme expliqué dans la partie 4.1.3.

Listing 2 – Unités du fichier .aa lu par Glozz

```
<unit id="TXT_IMPORTER_1594195313472270">
  <metadata>
    <author>TXT_IMPORTER</author>
    <creation-date>1594195313472270</creation-date>
    <lastModifier>n/a</lastModifier>
    <lastModificationDate>0</lastModificationDate>
  </metadata>
  <characterisation>
    <type>paragraph</type>
    <featureSet/>
  </characterisation>
  <positioning>
    <start>
      <singlePosition index="5"/>
    </start>
    <end>
      <singlePosition index="55"/>
    </end>
  </positioning>
</unit>
<unit id="adecker_1594195313472295">
  <metadata>
    <author>adecker</author>
    <creation-date>1594195313472295</creation-date>
    <lastModifier>n/a</lastModifier>
    <lastModificationDate>0</lastModificationDate>
  </metadata>
  <characterisation>
    <type>Segment2</type>
    <featureSet>
      <feature name="Nom">G1</feature>
    </featureSet>
  </characterisation>
  <positioning>
    <start>
      <singlePosition index="9"/>
    </start>
    <end>
      <singlePosition index="55"/>
    </end>
  </positioning>
</unit>
```

Utilisation des scripts

J'ai réalisé un troisième script permettant de centraliser les données : il suffit de compléter les différents champs tels que le fichier de travail et la liste d'interlocuteurs avant d'exécuter le script. Celui-ci exécutera les deux autres et créera donc trois fichiers : le texte numéroté, le fichier contenant les annotations Glozz et un autre fichier similaire mais avec seulement les unités permettant le découpage du texte.

4.1.3 Scripts python pour l'étude des annotations

L'ensemble des scripts représente environ 800 lignes de codes en plus d'un Jupyter Notebook comportant une centaine de lignes de codes.

Une partie des scripts permet de générer des graphes à partir des annotations Glozz. Ces graphes permettent d'étudier facilement un dialogue annoté car les segments sont placés horizontalement ou verticalement les uns par rapport aux autres en fonction des relations qui les lient. Cela permet par exemple de repérer les ruptures de la frontière droite.

Ces scripts avaient recours à une structure obsolète, j'ai donc dû les modifier pour qu'ils fonctionnent à nouveau.

Le Jupyter Notebook permet de créer divers graphiques mesurant le degré d'accord de plusieurs annotateurs sur un même texte : il a été réalisé dans le cadre des campagnes d'annotations. J'ai étudié son contenu lors de la phase bibliographique de mon travail mais ayant annoté de nouvelles transcriptions, je n'avais pas d'autres annotations avec lesquelles comparer les miennes, je n'ai donc pas utilisé ce Notebook. Cependant, j'ai modifié quelques lignes afin qu'il fonctionne à nouveau avec les modifications des scripts que j'avais faites.

Modifications réalisées

La structure *Panel* de la bibliothèque Pandas était utilisée pour représenter les relations entre les segments. Or cette structure est obsolète, le code ne fonctionnait donc plus.

Je l'ai remplacée par des *DataFrame*, comme ils n'ont que deux dimensions, j'ai utilisé la structure *MultiIndex* de la bibliothèque Pandas pour ajouter une dimension. J'ai ensuite modifié le code à plusieurs endroits, notamment les lignes qui manipulaient les *Panel* et changeaient leur forme, pour que le changement de structure ne pose pas de problème.

Conditions de fonctionnement des scripts de génération de graphe

Ces scripts nécessitent la lecture de trois fichiers : le fichier texte.ac qui contient tout le texte sans sauts de ligne, le fichier unites.aa qui est un fichier au format XML où la découpe en unités est signifiée et enfin un second fichier .aa qui contient toutes les relations.

Dans les fichiers .aa, les unités doivent comprendre un élément <featureSet> qui lui-même contient un élément <feature> Nom dont le texte est l'identifiant de l'unité (R12_3 par exemple) pour que les scripts fonctionnent.

4.2 Annotation des entretiens segment par segment

Après avoir pris en main les différents logiciels, j'ai travaillé sur l'annotation des deux entretiens en commençant par Ha-R. J'ai tout d'abord étudié seulement les paires question-réponse afin de déterminer si l'enchaînement des questions et des réponses posait des problèmes mais aussi pour avoir une idée globale du contenu de l'entretien. Ceci a permis de repérer une portion de la conversation qui semblait moins claire que le reste, je l'ai donc annotée complètement pour essayer de comprendre les raisons de cette impression. Ces premières annotations ont mis en évidence un manque dans la liste de relations utilisables pour les annotations. En effet, l'extrait étudié est plus long que ceux du corpus précédent, ce qui fait que de nouvelles relations entre les segments apparaissent comme lors des changements de sujets. C'est pourquoi j'ai complété la liste d'annotations avant d'annoter l'extrait de Ha-R mais aussi l'intégralité de Am-A, ce qui a permis de vérifier que la nouvelle liste était suffisamment complète.

4.2.1 Étude des paires question-réponse d'une conversation pathologique

La première étape était de repérer et annoter les différentes paires question/réponse dans la transcription Ha-R. Pour ce faire, j'ai utilisé le logiciel ELAN qui, en plus d'être utilisé pour transcrire les enregistrements, permet de créer toutes sortes d'annotations. Il y a environ 70 questions dans cette conversation, ce qui représente 10% des interventions. L'étude des paires question-réponse permet donc d'avoir un aperçu global du contenu de la discussion.

Travail sur l'entretien

Comme les entretiens sont transcris sur ELAN, j'ai complété directement le fichier de la transcription avec mes annotations. J'ai suivi le protocole d'annotations mis au point dans le cadre du projet tutoré (Cruz Blandón et al., 2018).

Le guide invite à créer cinq *tiers* pour les annotations : Question_Type, Answer_Type, Feature, Complexity et Is_Quoted. La Table 2 résume les annotations pouvant être utilisées pour chaque *tiers*.

J'ai ensuite parcouru la transcription et annoté les questions et les réponses en suivant la procédure conseillée par le guide.

Annotation des questions

Le premier *tiers* à compléter est Complexity. S'il y a une seule question, l'annotation est SQ (*single question*). Sinon, l'annotation est MQ (*multiple question*) et il faut subdiviser l'unité selon les sous-questions qu'elle contient comme sur la Figure 12 et suivre les étapes suivantes pour chaque sous-question.

Ensuite il faut compléter Question_Type en fonction du type de question.

Si la question était une question disjonctive (DQ) ou une Wh-question (WH) alors il faut compléter le *tiers* Feature avec l'annotation correspondant à la question. Certains

QUESTION_TYPE	YN (questions polaires) DQ (questions disjonctives) PQ (questions phatiques) CS (suggestion de complétion) WH (Wh questions)	Les seules réponses possibles sont oui et non. La question propose un choix entre plusieurs réponses. La question permet d'entretenir la conversation, elle n'a pas pour but d'obtenir une information. Un participant propose un mot ou un morceau de phrase pour terminer le propos d'un autre. Questions qui contiennent un mot interrogatif (comme Qui, Que, Quoi, Où, ...).
ANSWER_TYPE	PA (réponse positive) NA (réponse négative) FA (réponse avec fonction) PHA (réponse phatique) UA (réponse incertaine) UT (réponse hors sujet) DA (nie l'hypothèse)	Confirme la proposition contenue dans la question. Nie la proposition contenue dans la question. Réponse qui fournit l'élément demandé par la question (réponse aux Wh-Questions). Réponse sans contenu informationnel, a but purement communicatif. La personne n'est pas sure ou ne connait pas la réponse. La réponse n'est pas directement en relation avec la question, parfois elle y répond quand même. Contredit une présupposition de la question.
FEATURE	TMP (temporalité) LOC (lieux) AG (agent) TH (thème) OW (possesseur) RE (raison) CH (caractéristiques)	
COMPLEXITY	SQ (question simple) MQ (question multiple)	Il y a une seule questions dans la phrase. Il y a plusieurs question dans la phrase.
IS_QUOTED	QQ (question rapportée) NQ (question directe)	

TABLE 2 – Annotations possibles sur ELAN

mots interrogatifs tels que Qui, Où ou Pourquoi permettent de déterminer facilement l'annotation correcte mais souvent cette annotation est assez difficile à choisir.

Enfin il faut compléter le *tiers* IS_QUOTED selon que la question est au discours direct ou indirect.

Annotation des réponses

Le premier *tiers* à compléter est ANSWER_TYPE. Certains types de réponses sont plus probables que d'autres en fonction du type de la question.

Il faut ensuite compléter le *tiers* IS_QUOTED de même que pour les questions.

Résultats et analyse des annotations

En ce qui concerne le schéma d'annotation, certaines questions ne semblaient corres-

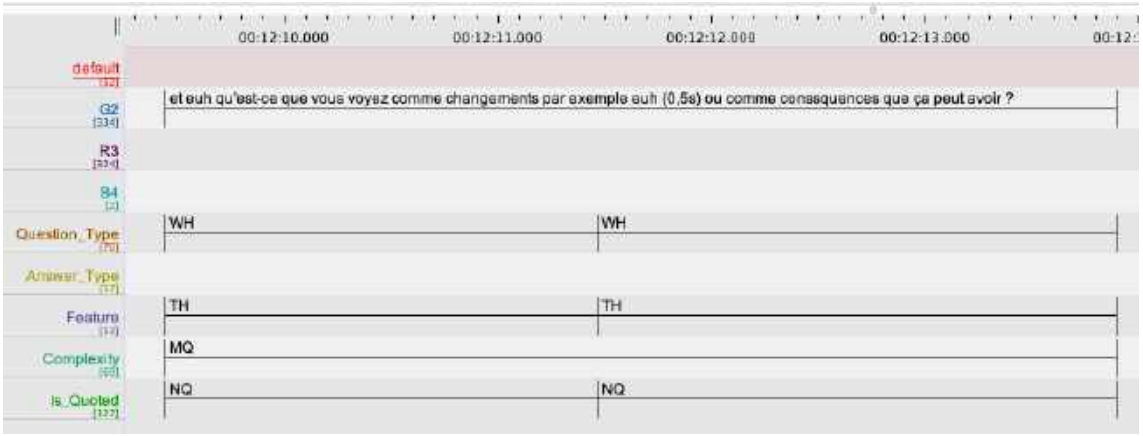


FIGURE 12 – Annotation d'une question multiple

pondre à aucun des types proposés. Il s'agit des questions visant à faire répéter l'autre locuteur ou demandant plus d'informations comme « Pardon ? », « De quoi ? » ou « Comment ça ? ». Ces questions semblent se situer entre des questions phatique et des Wh-questions mais aucune des deux catégorie n'est vraiment satisfaisante. Un nouveau type de question « Demande de clarification » (Purver, 2004) pourrait être envisagé. J'ai choisi par la suite d'utiliser Méta-Question, une des relations du schéma d'annotation Glozz, pour caractériser ces questions.

En se concentrant sur les questions et les réponses, aucun dysfonctionnement majeur n'apparaît : il y a très peu de questions sans réponses ou de réponses réellement hors sujet. Mais à la lecture de l'échange, on remarque qu'une partie de la discussion semble moins claire que le reste.

Dans la majeure partie du dialogue, le patient, très intéressé par l'objectif du projet auquel il prend part, pose des questions à ce sujet et partage se propre expérience des neurosciences. L'échange ne présente alors aucun signe de dysfonctionnement. Mais suite à une question du psychologue sur les raisons de son internement, la discussion s'éloigne du sujet et parait devenir plus chaotique avant de redevenir aussi claire qu'au début après une intervention du patient voulant à nouveau parler de ce qu'il appelle la « nouvelle science ».

4.2.2 Annotation précise d'un extrait de la transcription

Comme une partie de la discussion se distinguait, seul cet extrait a été annoté. L'objectif était de déterminer si la structure de cette partie de l'échange suivait les règles de la S-DRT afin de comprendre pourquoi elle paraissait plus étrange que le reste de la conversation.

Nouvelles relations discursives pour préciser les questions-réponses

La liste de relations retenue pour les campagnes d'annotations des deux projets tutorés (Huber & Laurier, 2017) et (Biver et al., 2018) ne comportait que les relations Question, Méta-Question et Réponse pour caractériser les questions et réponses. J'ai donc créé de

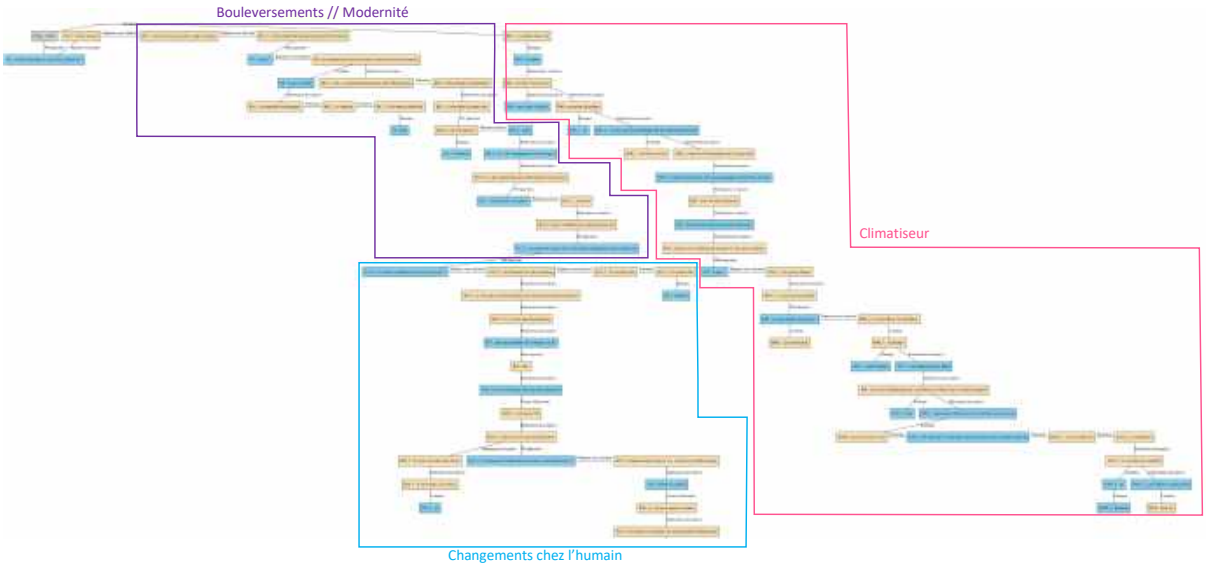


FIGURE 13 – Extrait du graphe d'annotations de Ha-R

nouvelles relations pour appliquer le même schéma d'annotations que sur ELAN.

J'ai créé une relation par type de question et réponses soit : Question disjonctive, Question phatique, Suggestion de complétion, Wh Question et YN Question ainsi que Nie l'hypothèse, Réponse phatique, Réponse avec fonction, Réponse incertaine, Réponse négative, Réponse positive et Réponse hors sujet.

Pour les question disjonctives et les Wh Question, j'ai ajouté un élément <feature> Fonction pouvant contenir les différentes fonctions listées dans le guide d'annotations ELAN.

Dans la mesure où le texte est segmenté avant d'être annoté, j'ai estimé que la complexité de la question ne nécessitait pas d'annotation : si une question est multiple elle sera découpée en plusieurs segments conversationnels.

Quant au *tiers* Is_Quoted, j'ai choisi de ne pas le reproduire dans le schéma d'annotations Glozz car je n'en avais pas l'utilité pour les textes que j'ai annotés. Mais si ce schéma devait être réutilisé pour des textes où cette précision serait utile, il est possible de rajouter un élément <feature> Is_Quoted à toutes les relations de type questions et réponse.

Analyse des annotations de la transcription

La FIGURE 13 est un extrait de la modélisation sous forme de graphe des annotations de la transcription HA-R. Cet extrait correspond au début de la conversation pour les thèmes Bouleversements // Modernité et Changements chez l'humain, quand au thème Climatiseur, il intervient vers la fin de la discussion après un bruit de climatiseur dans la pièce.

En ce qui concerne le modèle d'annotations, deux problèmes apparaissent ici. Tout d'abord la partie sur le climatiseur n'est pas évidente à rattacher dans la mesure où il s'agit d'une interruption de la discussion provoquée par un facteur extérieur, la discussion

R102_3 : le microbe il est amplifié
G103_1 : ah
G103_2 : humhum
G103_3 : ça le diffuse un peu partout
R104 : bien-sûr
G105 : **pour revenir à ce que vous disiez avant** vous avez une télé ?

FIGURE 14 – Extrait de l'échange Ha-R sur le climatiseur

reprenant son cours après en avoir terminé à ce sujet. Cette partie pose donc problème en termes de rattachement au début mais aussi à la fin : ici le bloc du climatiseur est rattaché au début de la discussion car il rompt avec ce qui est dit juste avant mais ceci constitue une rupture de la frontière droite pour la S-DRT alors que connaissant le contexte, cette partie de la discussion ne pose pas de problème. Pour la suite de la discussion, qui ici n'apparaît pas, elle peut se rattacher soit à la fin de la partie sur le climatiseur soit plus haut à la suite de la discussion plus générale. C'est ce second choix qui a été privilégié car la discussion reprend là où elle s'était interrompue lors du bruit de climatiseur. Cette remontée dans la structure ne pose pas de problème car comme le montre la fin de l'échange sur le climatiseur FIGURE 14, le thème est clos proprement et le retour à la conversation courante est évident. Cet échange amène donc à un questionnement sur les différents plans de conversation et le passage de l'un à l'autre ainsi que la façon de les représenter.

L'autre problème est lié au fait que les annotations sur le précédent corpus étaient faites sur des extraits assez courts. De ce fait, la liste de relations discursives, choisie pour être la plus simple possible, ne semble pas assez complète pour annoter un extrait aussi long. C'est pourquoi j'ai ré-annoté cet extrait avec une liste de relations plus complète.

4.2.3 Précision du modèle d'annotation

En plus de ré-annoter l'extrait précédent, j'ai annoté une conversation complète qui venait d'être transcrite. Ceci m'a permis de déterminer plus précisément les relations discursives manquantes.

Choix des nouvelles annotations

Comme les extraits annotés sont beaucoup plus longs qu'avec le précédent corpus, il y a plus d'introductions et de changements de sujets. C'est pourquoi j'ai ajouté une relation *Déclaration*. J'ai ajouté aussi *Déclaration incertaine* et *Méta-Déclaration* pour compléter les différents types de relations. Ce sont des relations coordonnantes donc horizontales.

Dans les deux extraits que j'ai annotés, les interlocuteurs avaient tendance à répéter ce que l'autre venait de dire, confirmant qu'il avait bien compris. Ces interventions sont proche des phatiques mais apportent quelque chose en plus dans la discussion, c'est pourquoi j'ai ajouté la relation *Confirmation*, subordonnante et donc verticale comme les phatiques.

J'ai aussi complété les relations de type question par la relation *Requête* puisque les requêtes ont souvent le même rôle qu'une question mais n'en n'ont pas nécessairement la

forme.

Enfin, j'avais supprimé la relation *Réponse* en pensant que les différentes relations plus précises seraient suffisantes. Mais en annotant les extraits, j'ai réalisé que lorsqu'un interlocuteur demande à l'autre de répéter quelque chose ou qu'ils se sont mal compris, la Méta-Question attend effectivement une réponse qui ne rentre dans aucune des catégories que j'avais ajoutées. J'ai donc remis cette relation dans la liste.

On se retrouve finalement avec les relations de la TABLE 3.

Narrations	Narration Déclaration Déclaration incertaine	**Méta**	Conduite Confirmation Phatique Méta-Déclaration Méta-Question Question phatique Réponse phatique
Élaborations	Élaboration descriptive Élaboration prescriptive Élaboration évaluative Contre-élaboration		
Questions	Question disjonctive YN Question Wh Question Suggestion de complétion Requête	826400extbfRéponses	Réponse Réponse avec fonction Réponse positive Réponse négative Nie l'hypothèse Réponse incertaine Réponse hors sujet

TABLE 3 – Ensemble final de relations discursives

Analyse des nouvelles annotations

La FIGURE 15 est un extrait de la modélisation sous forme de graphe des nouvelles annotations de la transcription HA-R. Cet extrait correspond à la deuxième moitié de la conversation. Dans l'ordre chronologique, les thèmes sont Smartphone, Climatiseur, TV, TV // Info-Manipulation puis Sarkozy. On peut observer que le graphe se divise en deux : le thème TV n'est pas rattaché juste après Climatiseur mais à la première unité du thème Smartphone.

Le questionnement à propos du climatiseur subsiste mais la relation Déclaration semble appropriée pour un rattachement linéaire du bloc climatiseur et il est possible de rattacher la suite de la discussion au précédent échange sur les smartphones et les télévisions sans problème lié à la remontée dans la structure comme expliqué précédemment.

Cette conversation ne pose donc pas de problème en terme de structure contrairement aux extraits du corpus précédent. En revanche, au niveau sémantique, l'échange semble très légèrement étrange, comme s'il n'était pas complet et que le passage d'un thème à l'autre n'était pas naturel.

Pour m'assurer que la nouvelle liste de relations rhétorique était suffisamment complète, j'ai annoté l'intégralité de l'extrait AM-A, soit environ 700 relations.

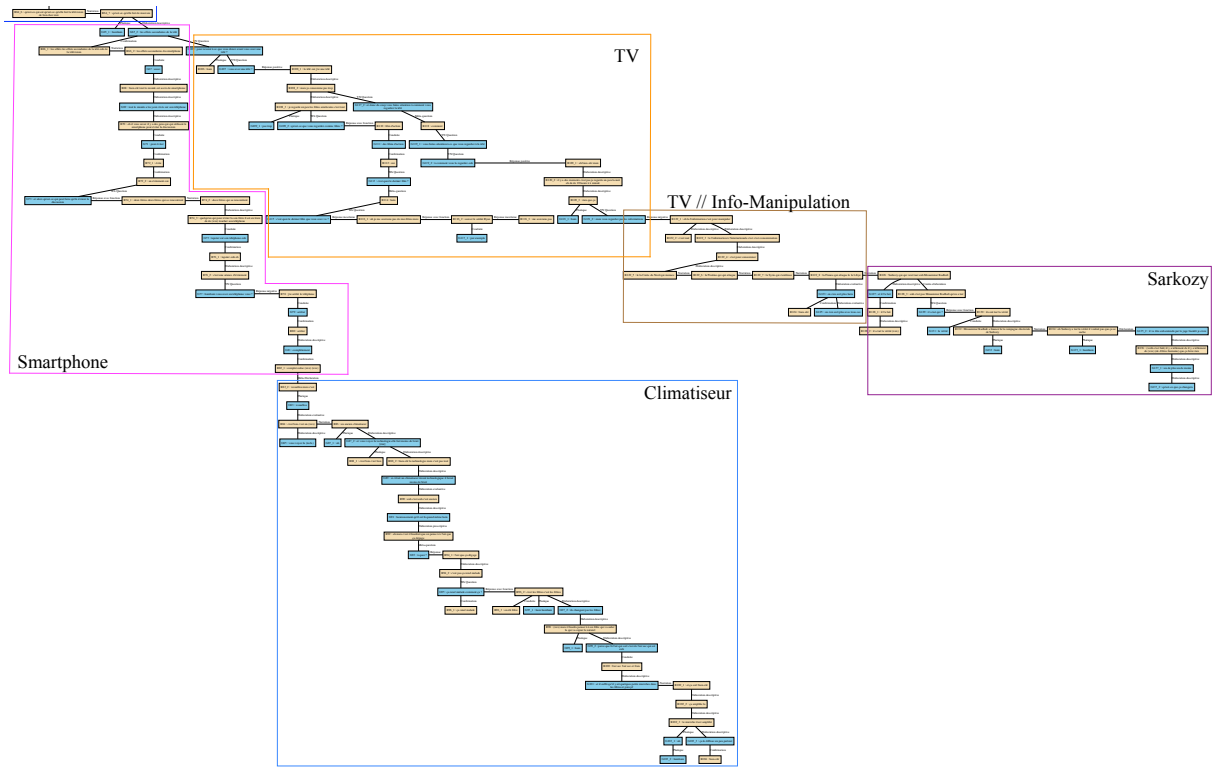


FIGURE 15 – Extrait du graphe d'annotations de Ha-R

La FIGURE 16 est un extrait de la représentation sous forme de graphe de ces annotations. Cet extrait se situe vers le milieu de la discussion et présente aussi une interruption extérieure : la porte s'ouvre et le patient cherche une explication pendant que le psychologue va la fermer.

Cet extrait ne présentait pas non plus de ruptures de la frontière droite ou alors à des endroits où d'autres rattachements sans rupture étaient possibles. En revanche, la présence du même genre d'interruption que celle du climatiseur permet donc d'apporter plus de matière à notre réflexion sur les différents plans de conversation.

4.3 Analyse thématique

La réflexion sur les plans de conversation est liée à l'analyse des boîtes thématiques. En effet, on ne change de plan que lorsqu'on change de thème dans les extraits que j'ai annotés. Mais on ne change pas de plan à chaque fois, j'ai donc cherché à comprendre comment les locuteurs passent d'un thème à un autre. Pour cela j'ai travaillé sur des annotations des boîtes thématiques. J'ai commencé par simplement expliquer ce passage grâce à leur contenu sémantique avant de réfléchir plutôt à la structure de la discussion de manière globale. J'ai travaillé sur HA-R pour mettre au point ce modèle d'annotation avant d'appliquer sa dernière version à AM-A.

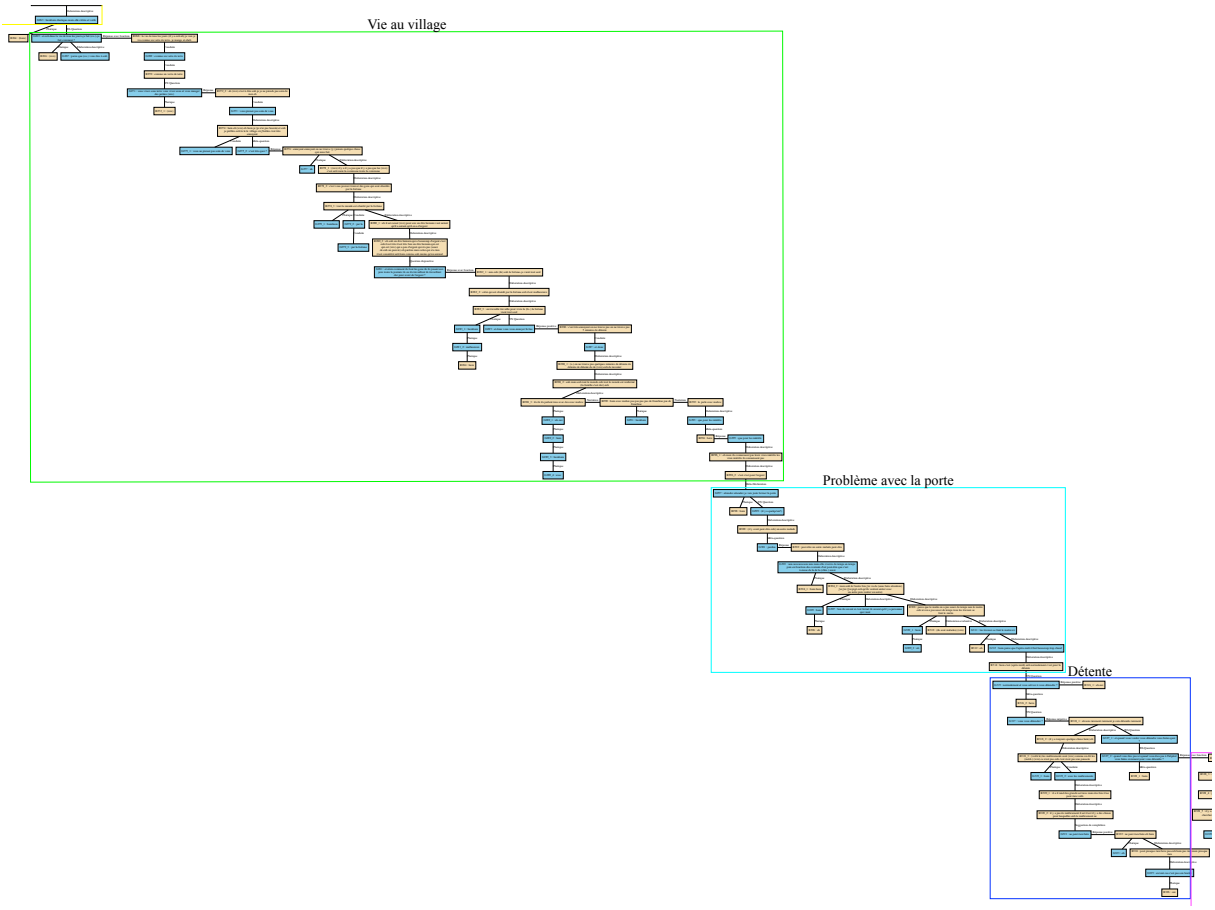


FIGURE 16 – Extrait du graphe d'annotations de Am-A

4.3.1 Première approche

Pour comprendre et modéliser l'enchaînement des idées dans la conversation, j'ai construit un premier schéma linéaire des boîtes thématiques de l'extrait de Ha-R que j'avais annoté. La Figure 17 reprend tous les thèmes et associe au passage de l'un à l'autre une explication liée au contenu sémantique du dialogue. Cette représentation permet de comprendre l'enchaînement des idées. Le thème Climatiseur y est en retrait puisqu'il intervient sur un autre plan de conversation.

Bouleversement // Modernité		
↓	La modernité et les bouleversements induisent des changements chez l'être humain	
Changements chez l'humain	- Un de ces changements est que l'humain ne va plus en profondeur	
	- La TV l'empêche d'aller en profondeur	
↓	- TV = organe de désinformation	
Désinformation		
↓	La désinformation / propagande a pour but d'endoctriner les peuples pour qu'ils fassent la guerre	
Guerre		
↓	À cause de toutes ces fautes informations, l'humain est devenu une machine	
Humain-Machine	humain = machine à cause de la technologie	
↓	Il faudrait mieux informer les gens sur la technologie	
Diffusion de l'info		
↓	Notamment sur (**les TV** et) les smartphones	
Smartphone		
→	Bruit ambiant	
		Climatiseur
	Reviennent à la conversation : il faudrait mieux informer les gens sur les effets secondaires de la TV	←
TV		
↓	À la TV les infos ont pour but de manipuler	
TV // Info-Manipulation	On nous surcharge en informations	
↓	Par exemple avec l'affaire Sarkozy-Kadhafi	
Sarkozy		

Figure 17 – Première modélisation des boîtes thématiques de l'extrait HaR

Cette représentation permet de mieux comprendre le passage d'un thème à l'autre mais est trop linéaire par rapport au contenu des boîtes : la conversation est en fait plus structurée. C'est pourquoi j'ai construit un second schéma, la Figure 18, où chaque boîte est reliée à une autre qui la précède mais pas nécessairement à la boîte juste antérieure. Ces liens sont étiquetés par une relation expliquant le rôle des boîtes les unes par rapport aux autres. Le passage d'une boîte à l'autre est de plus expliqué par un résumé du contenu

sémantique du dialogue.

Le contenu est le même que sur la FIGURE 17 mais la structure de la conversation est mise en avant, on a un aperçu du schéma argumentation du dialogue.

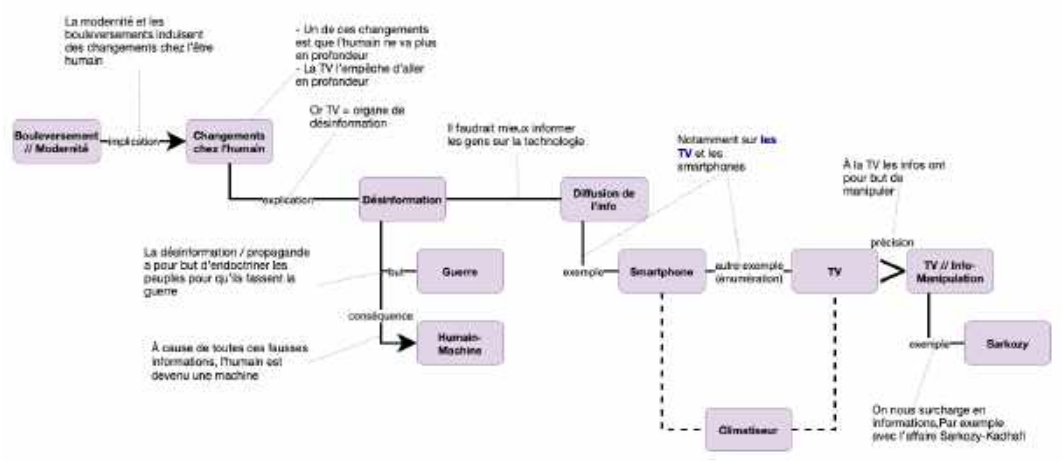


FIGURE 18 – Modélisation hiérarchisée des boîtes thématiques de l'extrait HaR

La boîte thématique Climatiseur est toujours à part, les pointillés signifient qu'elle intervient sur un autre plan de conversation que le reste du dialogue.

Cette version est plus hiérarchisée et permet de comprendre le cheminement de pensées liées à la discussion. La hiérarchie et les relations font penser à la S-DRT, c'est pourquoi j'ai remanié ce schéma pour en faire un graphe de type S-DRT.

4.3.2 Mise sous forme type S-DRT des analyses thématiques

Une représentation de type S-DRT implique d'assigner un type (coordonnante ou subordonnante) à chaque relation du schéma précédent. Dans les relations coordonnantes (horizontales), aucune partie ne domine l'autre tandis que dans les relations subordonnantes (verticales), un noyau domine un satellite.

Sur la FIGURE 19, le passage de la boîte 1 à la 2, de la 3 à la 6 ainsi que de la 7 à la 9 sont des continuations logiques de la conversation, ce qui explique le choix d'une relation coordonnante.

Pour les autres relations, une boîte domine toujours celle qui lui est rattachée. Les exemples des boîtes 7 et 11 précisent le contenu des boîtes 6 et 10 qui les dominent, la boîte 3 explique les phénomènes de la boîte 2 et les boîtes 4 et 5 découlent de la 3. C'est pourquoi des relations subordonnantes ont été choisies.

La relation liant la boîte du climatiseur à celle du smartphone est en pointillés pour signifier le changement de plan conversationnel. Une fois le sujet clos, la discussion reprend là où elle s'était arrêtée en passant à la boîte 9.

Les relations usuelles en S-DRT permettent de mettre en évidence la construction du discours. Les relations choisies pour la représentation des boîtes thématiques semblent plutôt expliquer le schéma argumentatif du dialogue. Ainsi il semblerait donc qu'il y a deux façons de comprendre un dialogue : d'une part une compréhension acte de dialogue

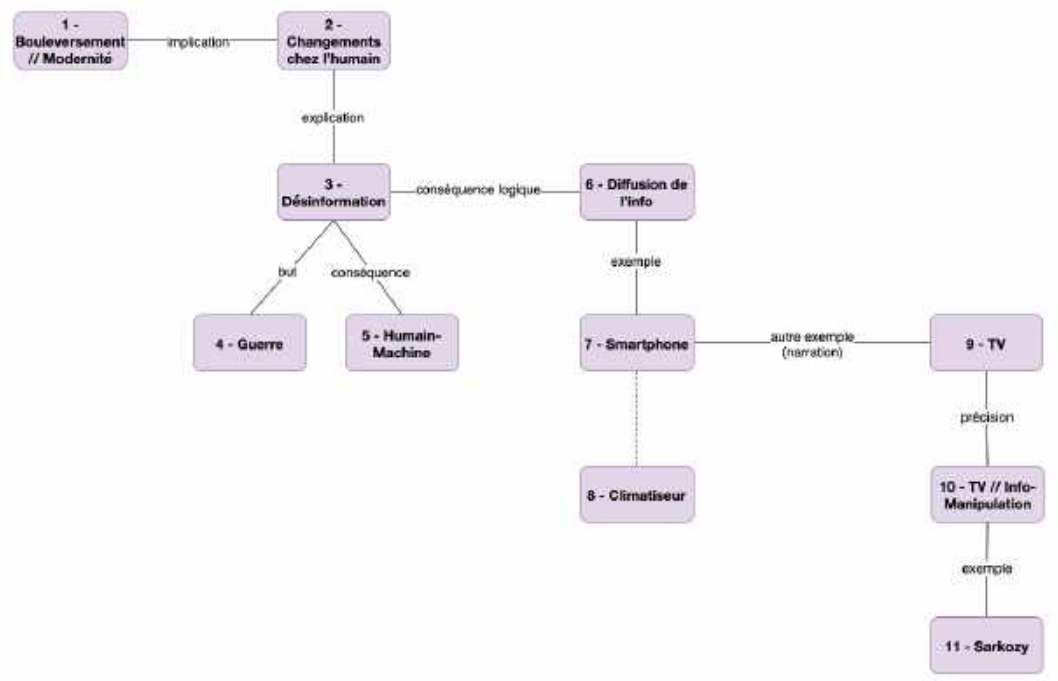


FIGURE 19 – Modélisation des boîtes thématiques sous forme S-DRT de l'extrait HaR

après acte de dialogue et d'autre part une compréhension plus globale qui permet de mettre en relation les différentes parties du dialogue.

Pour mieux observer la différence entre les deux types d'annotations, j'ai ajouté sur la FIGURE 20 un lien supplémentaire entre les boîtes. Ce lien reprend la même chronologie que les annotations Glozz et est annoté par la relation située au niveau du changement de thème.

On observe que les liens rouge ne suivent pas le même chemin que les noirs, ils sont plus linéaires, ce qui provient du fait que les annotations en S-DRT ont tendance à suivre le cours de la discussion. Le seul écart se situe au niveau du thème Climatiseur, ce qui confirme le statut différent de cette boîte par rapport à la conversation courante. La linéarité des liens rouges semble confirmer qu'il y a deux dynamiques dans le dialogue : l'enchaînement des actes de dialogue d'un côté et une structure globale où s'articulent les différentes parties du dialogue pour former un schéma argumentatif.

Pour étudier ce nouveau modèle d'annotation, je l'ai appliqué à l'intégralité de la transcription AM-A afin de pouvoir comparer les structures des deux conversations et voir si deux dynamiques apparaissent aussi dans cette conversation.

4.3.3 Comparaison des structures thématiques de deux conversations pathologiques

La structure de cette discussion semble assez différente au premier abord : dans toute la première moitié du dialogue, il ne s'agit pas d'un enchaînement de sujet mais plutôt d'un seul sujet majeur (le protocole) qui permet des digressions à plusieurs moments tout en revenant finalement au sujet.

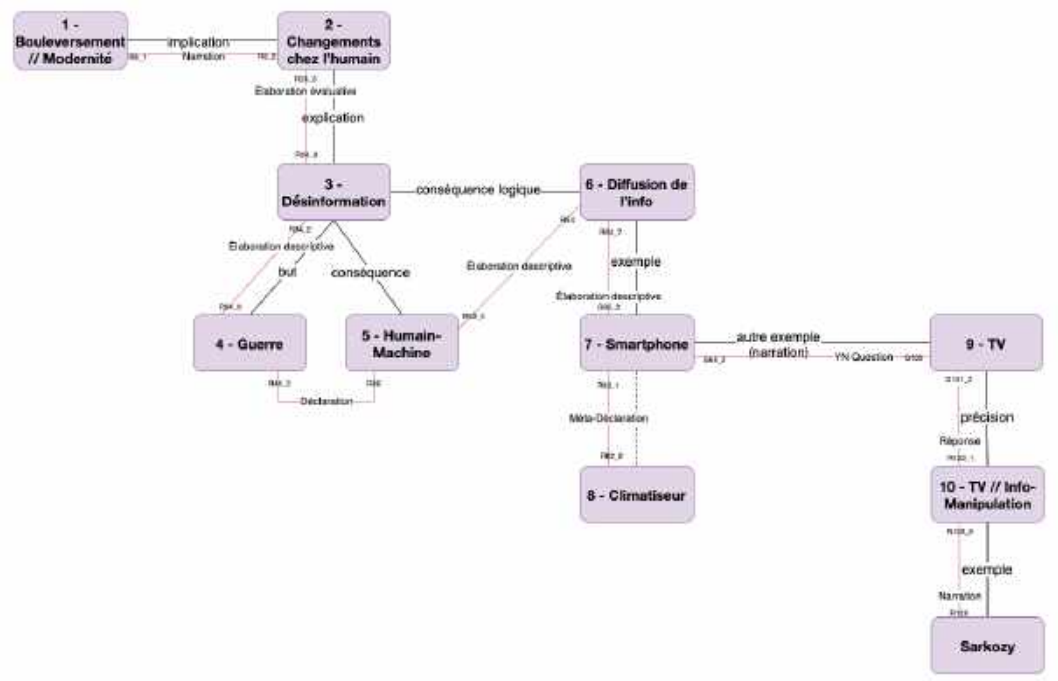


FIGURE 20 – Modélisation des boîtes thématiques sous forme S-DRT de l'extrait HaR avec annotations discursives

Ceci se voit sur la FIGURE 21 où jusqu'à la boîte 11, cinq boîtes au sujet similaire tournant autour du protocole sont alignées horizontalement sans se suivre chronologiquement. Elles dominent chacune une ou plusieurs autres boîtes alignées verticalement le cas échéant. Le lien rouge symbolisant l'avancée acte par acte est en revanche chronologique comme pour l'extrait précédent.

Les boîtes thématiques sont ensuite rattachées les unes aux autres de manière chronologique jusqu'à la boîte 16 où un événement extérieur fait que la conversation est interrompue et qu'une nouvelle boîte thématique apparaît : la porte s'est ouverte pendant la discussion et cela pousse les deux interlocuteurs à s'interroger sur les raisons pour lesquelles c'est arrivé. Cependant contrairement à l'extrait précédent, la discussion ne reprend ensuite pas où elle s'était arrêtée avant l'interruption, le psychologue se sert de la discussion autour de la porte pour entamer un nouveau sujet. Ceci est visible grâce au lien rouge qui ne présente pas d'interruption contrairement à celui de l'extrait HA-R. Cela montre qu'il y a plusieurs possibilités pour sortir d'un plan de conversation : ici on revient à la discussion courante tout en ouvrant une nouvelle boîte thématique sans continuité thématique avec la précédente (passage de Vie au village à Détente) alors que dans l'autre extrait la discussion revient explicitement au sujet précédent. L'intitulé du thème que j'ai choisi est différent mais le contenu est le même : si j'avais choisi des thèmes moins précis, tel que Technologie à la place de TV et Smartphone, la discussion aurait repris avec exactement le même thème qu'avant l'interruption.

Enfin cet extrait présente un troisième plan conversationnel qui apparaît avec la boîte 20. Il est créé par le patient lui-même sans interruption extérieure lorsqu'il coupe court à la discussion et expose son désir de s'en aller. La conversation se poursuit finalement encore un peu sans qu'elle ait de rapport avec ce qui a été dit avant mais il aurait été

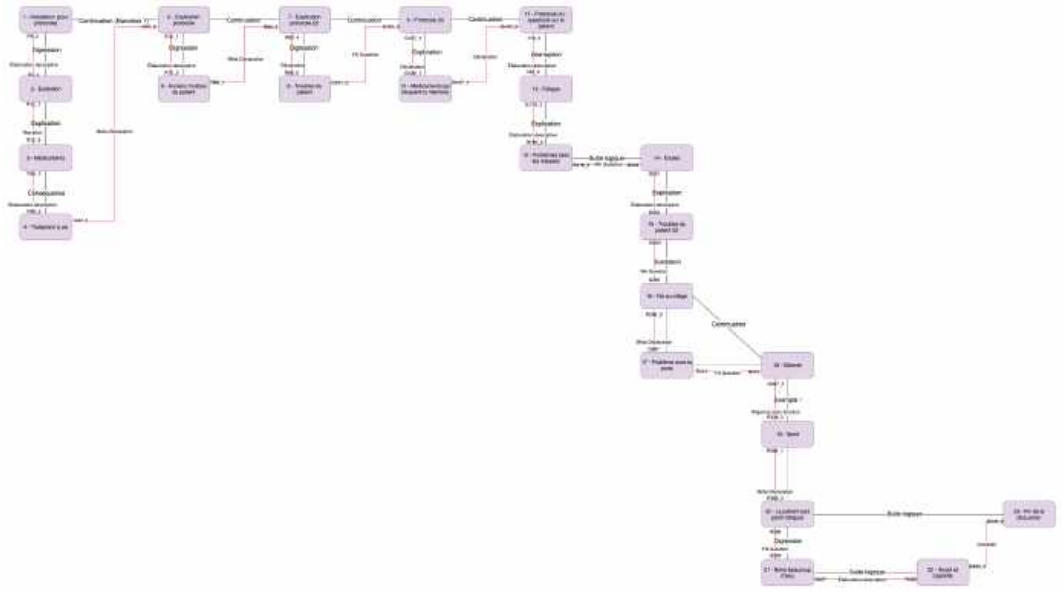


FIGURE 21 – Modélisation des boîtes thématiques sous forme S-DRT de l'extrait AmA avec annotations discursives

intéressant d'observer un retour au premier plan conversationnel pour mieux comprendre les possibilités qui existent pour passer d'un plan à un autre.

La modélisation thématique de ce second extrait confirme l'existence des deux dynamiques mise en avant par les liens noirs et les liens rouges. Il faut cependant prendre en compte le fait que le choix des thèmes et leur positionnement précis dans la discussion sont peut être moins consensuels que le choix des relations rhétoriques. En effet lors des campagnes d'annotation, le choix des thèmes que ce soit leur nombre ou leurs positions posait problème aux annotateurs. Cependant, les extraits alors annotés étaient bien plus courts que celui-ci, il est donc possible que la difficulté à choisir des thèmes adaptés provenait du fait que les textes étaient trop courts pour présenter différents sujets.

Finalement l'analyse des deux extraits n'a pas mis en évidence de ruptures ou d'incohérences mais elle a permis de remarquer une autre dynamique, de plus haut niveau, dans la conversation. Pour l'analyser, la S-DRT avait une granularité trop fine, ce qui a mené à la réflexion sur un modèle d'annotation plus abstrait, inspiré des principes de la S-DRT qui s'ajoute au modèle initial. Cette nouvelle représentation fait apparaître de nouveaux questionnements sur les plans conversationnels : les deux transcriptions modélisées ne permettent pas de dire comment s'effectue le passage d'un plan à l'autre même s'il semblerait que les relations rhétoriques de type Meta puissent jouer un rôle dans l'apparition de nouveaux plans. La combinaison des deux représentations, celle en S-DRT fine et celle des boîtes thématiques, pourrait aider à expliquer le fonctionnement des plans.

5 Conclusion et perspectives

Durant ce stage, j'ai pu effectuer un travail bibliographique concernant d'une part les publications liées au projet SLAM lui-même et d'autre part sur la sémantique et la

formalisation sémantico-discursive. J'ai ainsi eu un aperçu des difficultés et des particularités d'un projet pluridisciplinaire et j'ai pu en apprendre plus sur la façon de définir un modèle de représentation du discours en partant d'un formalisme plutôt complexe pour l'adapter aux besoins du projet. J'ai de plus acquis quelques bases en étude du discours et j'ai notamment pu voir l'évolution des théories à ce sujet de la Grammaire de Montague à la S-DRT, chacune cherchant à résoudre un problème non résolu par les théories précédentes.

Dans le cadre de la prise en main des différents outils utiles à l'annotation de texte, j'ai étudié les scripts python réalisés par des étudiants lors du projet tutoré (Biver et al., 2018). Il est toujours intéressant d'examiner le code de quelqu'un d'autre et ce travail était d'autant plus intéressant qu'il m'a fallu modifier une partie du script à cause d'une fonction obsolète depuis la mise à jour d'un module. J'ai aussi rédigé quelques scripts python moi-même pour automatiser la préparation des document avant le travail d'annotation sur Glozz. J'ai donc pu améliorer mes aptitudes en python.

En travaillant sur deux des dernières transcriptions de conversations pathologiques, j'ai pu apporter une première version d'annotations en prenant en compte les travaux sur les questions et réponses qui ont été produits depuis l'analyse du précédent corpus. Il s'agissait de plus d'extraits bien plus longs que ceux annotés dans le précédent corpus, ce qui a permis de réfléchir à de nouvelles relations discursives afin de préciser les annotations.

Ces extraits étant bien plus longs, les thèmes abordés au cours des discussions étaient plus nombreux, ce qui a permis de mettre en évidence un autre schéma d'annotation permettant d'expliquer l'articulation des boîtes thématiques les unes par rapport aux autres. Cette nouvelle vision a aussi mis en évidence la présence de plusieurs plans conversationnels liés notamment aux méta-interventions qui ont tendance à créer un nouveau plan.

Il serait donc intéressant de pousser la réflexion sur ces plans conversationnels et notamment les règles permettant le passage de l'un à l'autre. Reproduire le travail d'annotations thématiques avec d'autres transcriptions pourrait être intéressant non seulement pour définir la portée et la signification de ce schéma mais aussi pour éventuellement définir les règles qui le régissent. Comme il est inspiré de la S-DRT, on peut imaginer qu'une règle semblable à celle de la frontière droite existe. De plus, en superposant les annotations en S-DRT aux annotations thématique sur plus de conversations, il est possible que certaines relations rhétoriques s'avèrent avoir une influence particulière au niveau des thèmes, certaines seraient par exemple plus souvent à l'origine de la création d'un nouveau thème, ou d'un nouveau plan que d'autres.

Enfin, les thèmes ne faisaient pas consensus lors des campagnes d'annotations. Il n'est pas impossible que le désaccord fut lié au fait que les extraits analysés étaient courts et présentaient donc peu de thèmes facilement discernables mais il serait intéressant de vérifier ceci avec une campagne d'annotation centrée sur la définition des thèmes.

Références

Amblard, M., Boritchev, M., Carletti, M., Dieudonat, L., & Tsai, Y. (2019a). A Taxonomy of Real-Life Questions and Answers in Dialogue. In *SemDial 2019 - LondonLogue - 23rd Workshop on the semantics and pragmatics of dialogue*. London, United Kingdom.
URL https://hal.inria.fr/hal-02269609

Amblard, M., Musiol, M., & Rebuschi, M. (2011). Une analyse basée sur la S-DRT pour la modélisation de dialogues pathologiques. In M. Lafourcade, & V. Prince (Eds.) *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles - TALN 2011*, (pp. 93 – 98). Montpellier, France : Laboratoire d'Informatique de Robotique et de Microélectronique.
URL https://hal.archives-ouvertes.fr/hal-00601622

Amblard, M., Musiol, M., & Rebuschi, M. (2012). Schizophrénie et Langage : Analyse et modélisation. De l'utilisation des modèles formels en pragmatique pour la modélisation de discours pathologiques. In *Congrès MSH 2012*. Caen, France.
URL https://hal.archives-ouvertes.fr/hal-00761540

Amblard, M., Musiol, M., & Rebuschi, M. (2014). L'interaction conversationnelle à l'épreuve du handicap schizophrénique. *Recherches sur la philosophie et le langage*, *31*, 1–21.
URL https://hal.archives-ouvertes.fr/hal-00955660

Amblard, M., Rebuschi, M., & Musiol, M. (2019b). Corpus et pathologie mentale : particularités dans la constitution et l'analyse d'une ressource. In M. Rebuschi, & C. Benzitoun (Eds.) *Les corpus en sciences humaines et sociales*. Presses Universitaires de Nancy.
URL https://hal.inria.fr/hal-02269622

Asher, N., Asher, N., Lascarides, A., Press, C. U., Bird, S., Boguraev, B., Hindle, D., Kay, M., McDonald, D., & Uszkoreit, H. (2003). *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.
URL https://books.google.fr/books?id=VD-8yisFhBwC

Barwise, J., & Moravcsik, J. (2014). Formal philosophy. selected papers of richard montague. edited and with an introduction by thomason richmond h.. yale university press, new haven and london 1974, 369 pp. *The Journal of Symbolic Logic*, *47*, 210–215.

Biver, A., Colin, E., & Lefebvre, C. (2018). Trouble du langage et de la pensée : Campagne d'annotation.

Boritchev, M., & Amblard, M. (2019). Picturing Questions and Answers - a formal approach to SLAM. In M. Amblard, M. Musiol, & M. Rebuschi (Eds.) *(In)coherence of discourse - Formal and Conceptual issues of Language*. Springer. Language, Cognition and Mind.
URL https://hal.inria.fr/hal-02269631

Busquets, J., Vieu, L., & Asher, N. (2001). LA SDRT : Une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum (Presses Universitaires*

de Nancy), *13*(1), 73–101.
URL https://hal.archives-ouvertes.fr/hal-01686256

Carletti, M., Dieudonat, L., & Tsai, Y. (2019). Where's the answer : Dialogue annotation.

Cruz Blandón, M. A., Minnema, G., & Nourbakhsh, A. (2018). What's the answer : Dialogue annotation.

Geurts, B., Beaver, D. I., & Maier, E. (2020). Discourse Representation Theory. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Spring 2020 Edition.
URL https://plato.stanford.edu/archives/spr2020/entries/discourse-representation-theory/

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.) *Syntax and Semantics*, vol. 3, (p. 45–47). New York : Academic Press.

Heim, I. (2008). *File Change Semantics and the Familiarity Theory of Definiteness*, (pp. 223 – 248).

Huber, L., & Laurier, E. (2017). Trouble du langage et de la pensée : Campagne d'annotation.

Kamp, H., & Reyle, U. (1993). From discourse to logic : Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory.

Lascarides, A., & Asher, N. (2007). *Segmented Discourse Representation Theory : Dynamic Semantics With Discourse Structure*, vol. 3, (pp. 87–124). Kluwer.

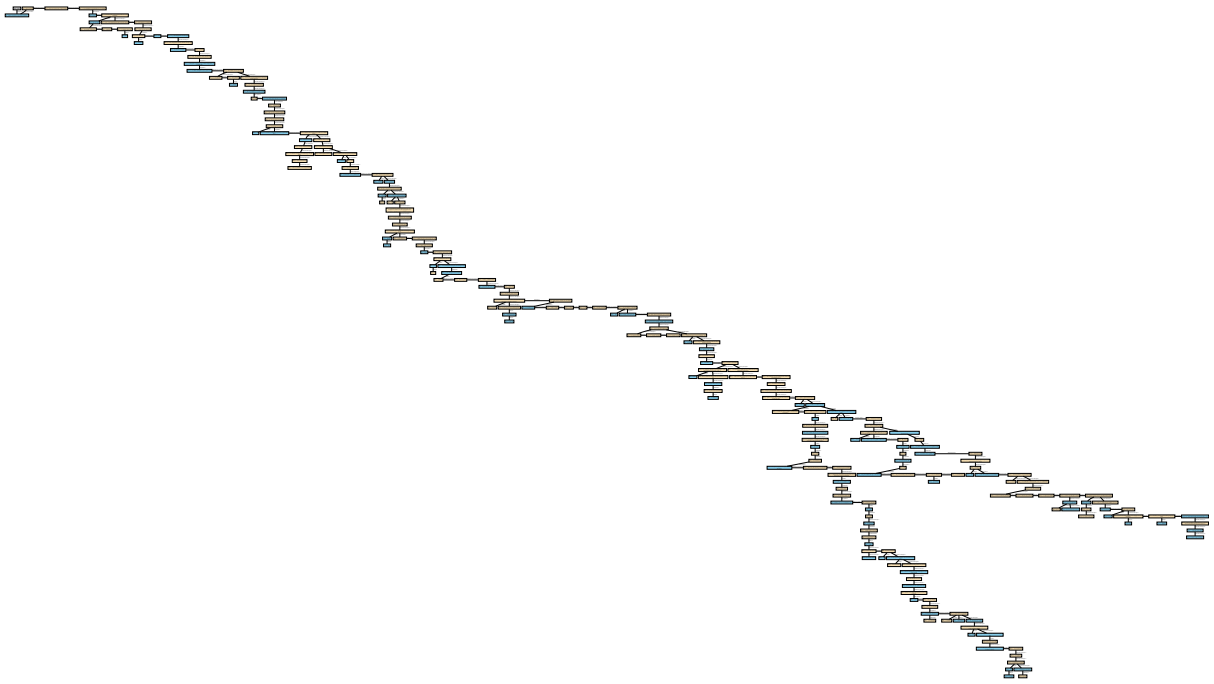
Purver, M. (2004). *The Theory and Use of Clarification Requests in Dialogue*. Ph.D. thesis, Department of Computer Science, King's College, University of London.

Rebuschi, M., Amblard, M., & Musiol, M. (2013). Schizophrénie, logicité et compréhension en première personne. *L'Évolution Psychiatrique*, *78*(1), 127–141.
URL https://hal.archives-ouvertes.fr/hal-00869681

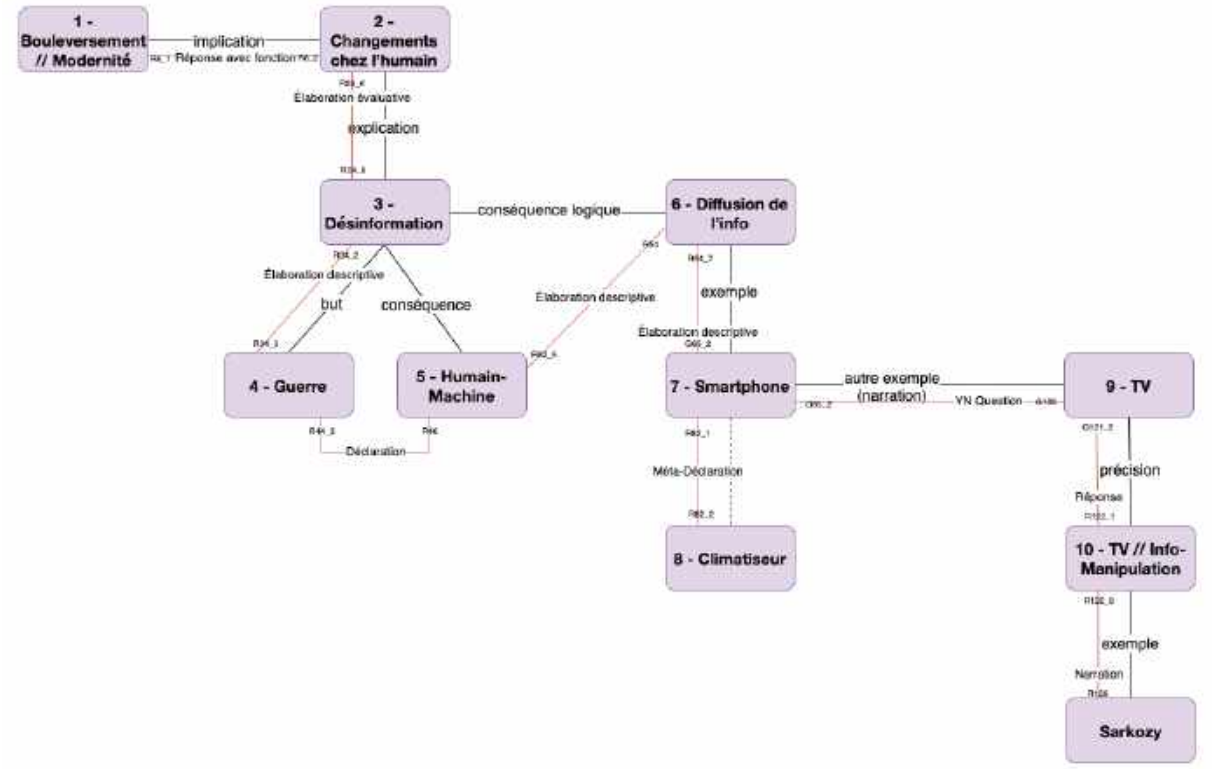
Schlöder, J. (2019). Discourse structure in dialogue.

A Travail sur Ha-R

A.1 Annotations Glozz

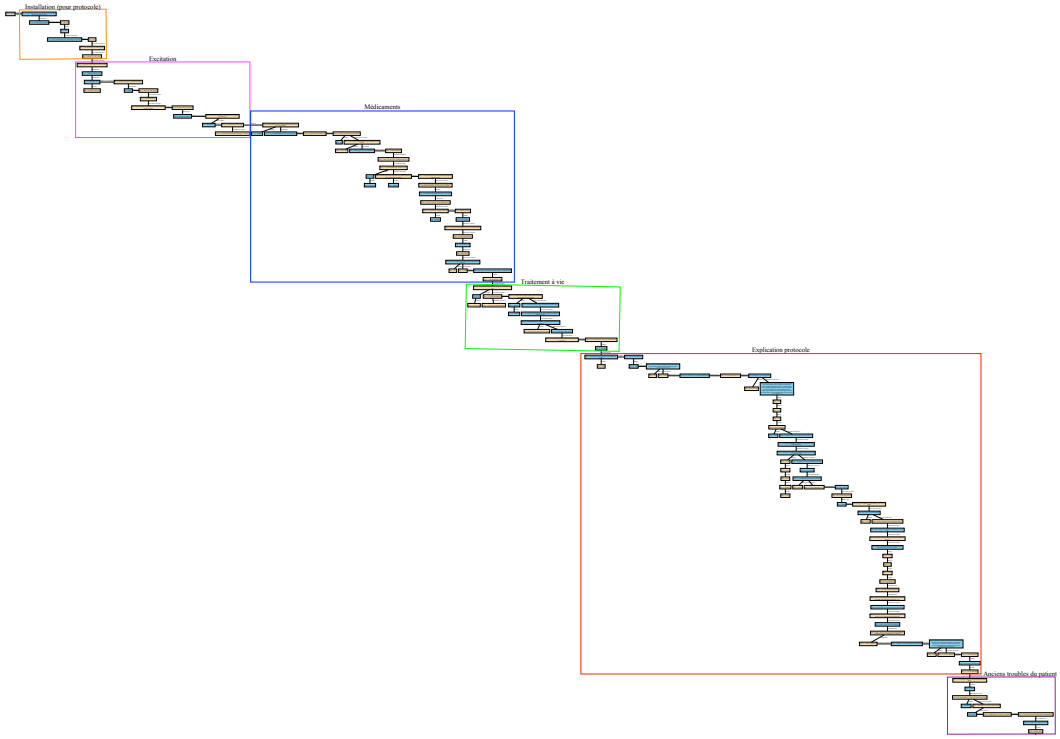


A.2 Annotations thématiques

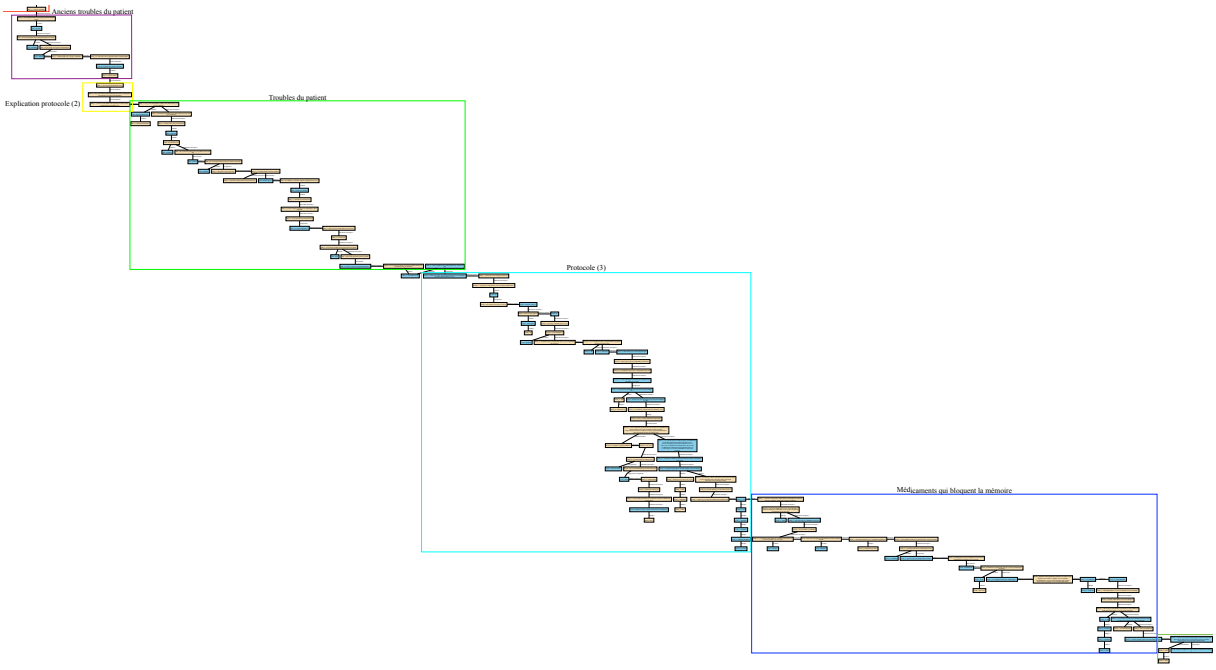


B Travail sur Am-A

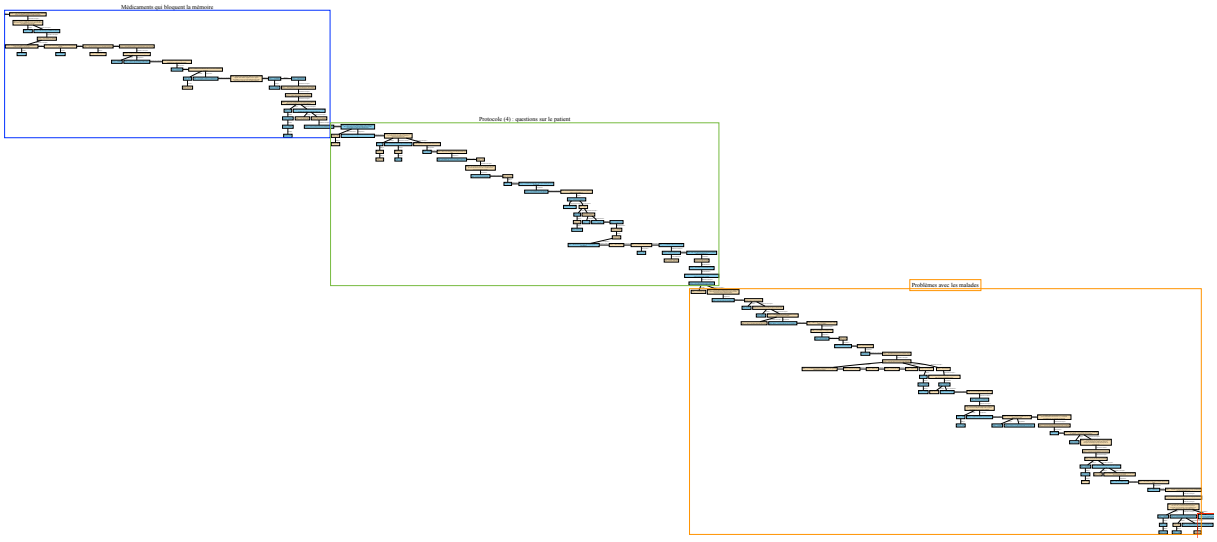
B.1 Annotations Glozz



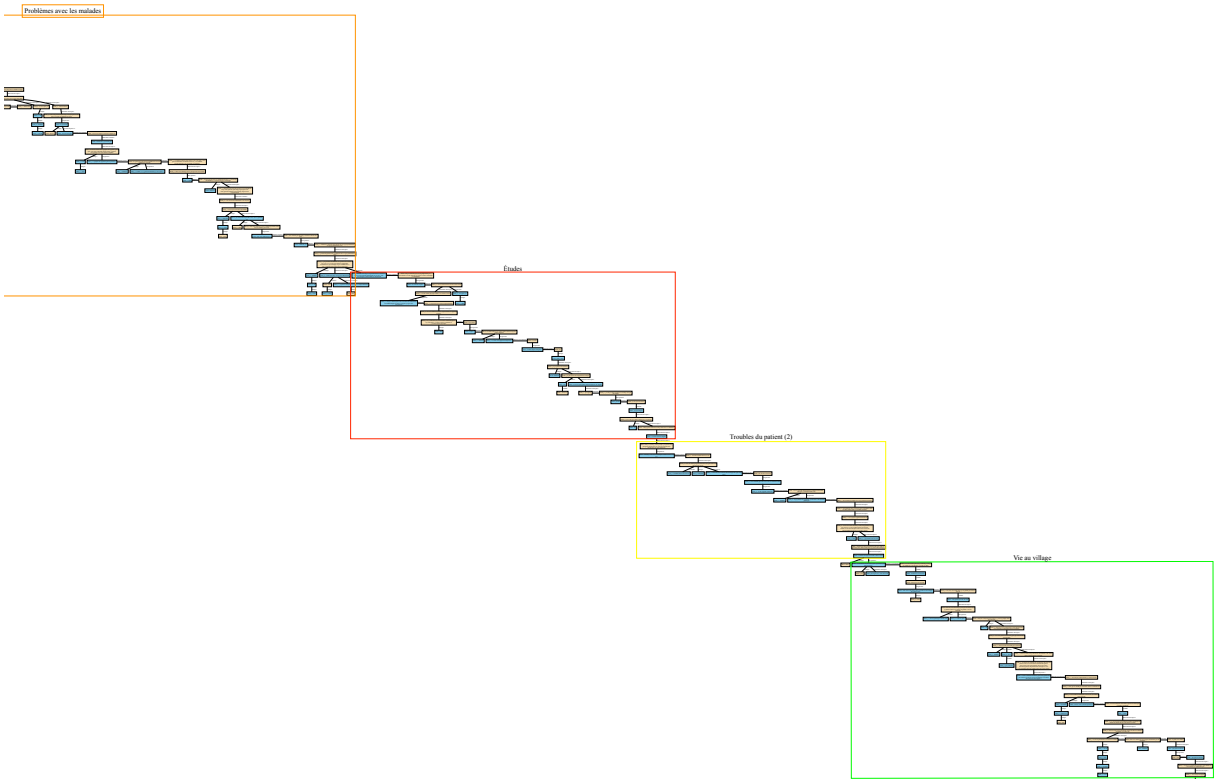
Partie 1/5



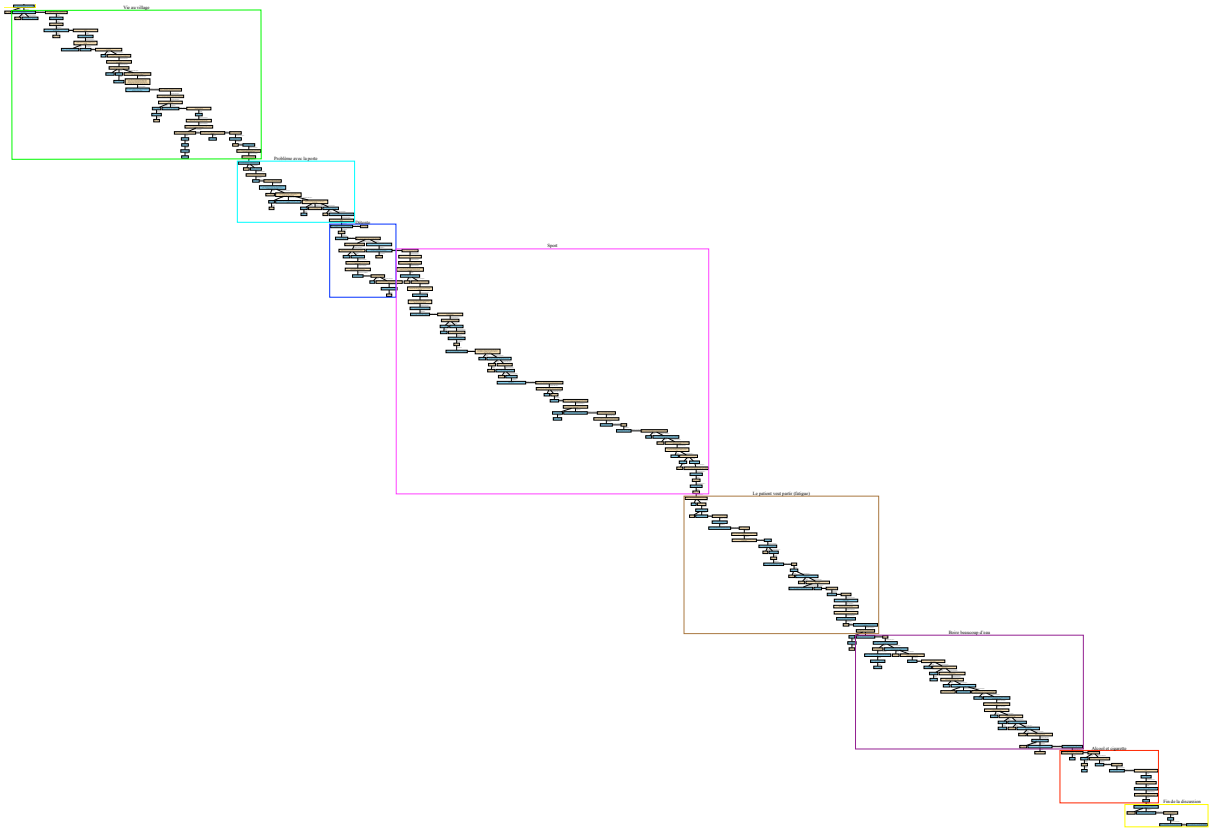
Partie 2/5



Partie 3/5



Partie 4/5



Partie 5/5

B.2 Annotations thématiques