Projet tutoré

Troubles du langage et de la pensée : campagne d'annotation

Etudiants :

Alexis Biver

Emile Colin

Corentin Lefebvre

Tuteurs :

Maxime Amblard

Manuel Rebuschi

Master 1 Sciences de la cognition et applications - 2017/2018

Projet tutoré

Troubles du langage et de la pensée : campagne d'annotation

Etudiants :

Alexis Biver

Emile Colin

Corentin Lefebvre

Tuteurs :

Maxime Amblard

Manuel Rebuschi

Master 1 Sciences de la cognition et applications - 2017/2018

REMERCIEMENTS

Merci à Maxime Amblard et Manuel Rebuschi pour nous avoir guidé tout au long du projet en nous apportant des retours pertinents sur notre travail et en nous aidant à le structurer afin de construire un plan et un planning essentiel au bon déroulement du projet tutoré.

Merci à tout le personnel administratif de l'université de Lorraine de nous avoir apporté du matériel, comme le prêt d'ordinateurs, pour mener au mieux notre campagne d'annotation. Et merci à Yann Mathet et Antoine Widlochet pour nous avoir fourni les identifiants Glozz nécessaires à son utilisation complète, facilitant ainsi grandement notre travail.

Merci à nos prédécesseures Laurine Huber et Emilie Laurier, pour leur travail réalisé dans ce projet de recherche qui a facilité de manière absolument considérable notre avancée. Merci à elles également d'avoir été disponibles pour répondre à nos questions et pour nous faire part des livrables dont nous avions besoin.

Merci à tous nos annotateurs, proches, connaissances ou inconnus mais curieux et intéressés, d'avoir bien voulu prendre le temps de passer ces annotations. Nous savons que cela représente un travail relativement long et demande une réflexion importante. Merci à chacun d'être parvenu jusqu'au bout de la tâche en effectuant des annotations sérieuses.

TABLE DES MATIERES

1 INTRODUCTION

Dans le cadre de notre première année de Master Sciences Cognitives et Applications, nous avons effectué un projet tutoré au sein du projet SLAM sous la supervision de Manuel Rebuschi (AHP-PReST)[1] et Maxime Amblard (LORIA)[2]. Pendant tout un semestre, notre mission a été de mettre en œuvre, dans la suite des travaux réalisés précédemment, une campagne d'annotation, auprès du grand public, de textes issus de conversations avec des personnes schizophrènes, puis d'analyser les données recueillies.

Ce rapport a pour but de présenter le contexte dans lequel s'inscrit le projet SLAM et l'expérimentation que nous avons mené, le déroulement de cette expérience ainsi que les résultats obtenus, et de porter un regard critique sur le travail que nous avons accompli et notre expérience personnelle. Nous présenterons donc dans une première partie le cadre de ce projet, c'est-à-dire le contexte institutionnel et scientifique dans lequel il s'inscrit, ainsi que ce qui était attendu de nous. Nous détaillerons ensuite le travail que nous avons effectué en amont pour nous approprier le sujet et les travaux précédents, puis la manière dont nous avons mis en place la campagne. Enfin nous présenterons les résultats que nous avons obtenus grâce à cette campagne.

1.1 CONTEXTE INSTITUTIONNEL

Le projet SLAM[3] (Schizophrénie et Langage : Analyse et Modélisation) est un projet de recherche interdisciplinaire porté par la Maison des Sciences de l'Homme (MSH) de Lorraine et par le Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) depuis 2013. Ce projet conjugue des approches de type informatique, philosophique, linguistique et psychologique pour analyser des conversations impliquant des personnes atteintes de schizophrénie et tenter de formaliser et théoriser les discontinuités caractéristiques de ce type de discours.

1.2 CONTEXTE SCIENTIFIQUE

Le sujet d'étude porte ici sur les troubles du langage et de la pensée et notamment chez les schizophrènes et leur rapport aux propriétés du langage, du discours et de la conversation. Ce projet voudrait définir le plus précisément possible les processus de pensée chez les schizophrènes.

Pour cela, il faut se baser sur un principe philosophique nommé le principe de charité et qui consiste à considérer que le comportement d'autrui est rationnel, en lui attribuant une intelligence et des facultés élevées. Ce principe est utilisé pour essayer de montrer qu'il n'y a pas de contradiction dans le discours des schizophrènes, mais que la rupture qu'il peut y avoir dans leur discours tient surtout d'une mauvaise utilisation des règles de la conversation.

Dans la pratique, nous faisons annoter des dialogues retranscrits entre des patients et un psychologue par des annotateurs dits naïfs (ne cherchant pas à analyser des phénomènes de rupture sémantique et pragmatique), et l'analyse de ces annotations permettra de confirmer ou d'infirmer l'hypothèse selon laquelle la logique interne des schizophrènes est préservée.

[1] Archives Henri-Poincaré - Philosophie et Recherches sur les Sciences et les Technologies

[2] Laboratoire Lorrain de Recherche en Informatique et ses Applications

[3] http://www.msh-lorraine.fr/index.php?id=626

1.3 OBJECTIFS DU PROJET

Notre objectif en nous intégrant à ce projet est de mener une nouvelle campagne pour récolter plus de données d'annotateurs dits naïfs. Nos prédécesseures ont déjà réalisé un travail conséquent d'affinage de préparation de la campagne ; nous avons donc reproduit la même méthodologie qu'elles, à ceci près que nous avons ajouté au panel de textes présentés un texte à rupture supplémentaire et deux nouveaux textes témoins, pour soumettre les annotations de dialogues présentant des ruptures à comparaison (tous les textes sont disponibles en Annexe 2).

Dans la pratique, la campagne s'est déroulée comme suit :

- Mener une précampagne : afin de s'adapter à la plateforme Glozz, nous avons dû réaliser un travail en amont pour nous former en tant que bons expérimentateurs, et modifier les critères choisis par rapport à la précédente campagne.
- Recruter des annotateurs : par le biais d'évènements, de communication via les réseaux sociaux, la réservation de lieux d'annotations, et la prise de rendez-vous; tout en essayant d'homogénéiser les candidats ciblés.
- Faire passer les annotations : Expliquer les consignes en présentiel aux annotateurs, rester présent en cas de problème, enregistrer les résultats.

Enfin, la majeure partie du travail qu'il nous a été confié se trouve dans l'analyse en profondeur des données récoltées de notre campagne et de la précédente. En effet, nos camarades de l'année précédente, en raison de leur travail sur la procédure, n'ont pas eu le temps d'analyser absolument toutes les données. Nous avons donc formulé de nouvelles hypothèses en ayant mis l'accent sur la recherche de consensus dans les annotations à plusieurs niveaux et la détection des ruptures discursives.

1.3.1 Étapes du projet

Dans la pratique, nous avons opéré chronologiquement selon le schéma suivant :

- Nous avons tout d'abord réalisé un travail bibliographique de longue haleine, nous permettant de comprendre ce sujet très complexe que représente l'analyse et la modélisation du discours dans le contexte des troubles du langage et de la pensée. De plus, les prémices de ce sujet sont étudiées depuis de nombreuses années [3].
- Ensuite, nous avons lu et repris le travail de Laurine Huber et Emilie Laurier [2], chargées du même projet tutoré pour l'année 2016-2017, afin de poursuivre un travail à long terme de manière coordonnée.
- À partir de cela, nous avons défini un calendrier (disponible en Annexe 1) sous la forme d'un diagramme de Gantt, pour toutes les tâches que nous devions réaliser au cours du projet.
- Dès lors et jusqu'au rendu du rapport, nous avons travaillé petit à petit sur l'analyse des données. Celle-ci exigeait du travail à long terme, étant donné que nous avions eu quelques difficultés à reprendre intégralement les données résultantes des analyses précédentes.
- L'analyse des données s'est déroulée parallèlement à la préparation de la campagne et le déroulement de cette dernière.
- Enfin, quand nous avons obtenu les résultats de la campagne, nous avons pu commencer à interpréter les données dans le but de répondre à nos hypothèses initiales.

2 TRAVAIL PRÉLIMINAIRE

2.1 TRAVAIL BIBLIOGRAPHIQUE

Durant notre travail bibliographique, nous nous sommes familiarisés avec les travaux traitant de l'analyse des discours des patients schizophrènes.

Comme mentionné dans le paragraphe 1.2, nous appliquons le principe de charité au discours d'un patient schizophrène, c'est à dire que nous considérons que ce discours est cohérent. S'il semble comporter des ruptures, elles ne peuvent donc pas être de nature sémantique (ce qui traduirait des pensées contradictoires chez le schizophrène), mais sont de nature pragmatique (le schizophrène ne respecte pas certaines règles du discours).

On identifie 2 types de ruptures discursives : la rupture de la frontière droite qui est un rattachement interdit à une unité du dialogue qui n'est plus accessible au moment de son rattachement; et le débrayage conversationnel, c'est-à-dire l'ouverture d'un nouvel arbre de conversation sans avoir bien fermé le précédent.

Nous nous sommes aussi rapidement formés à l'utilisation des principes de l'approche pragmatique et sémantique de la Segmented Discourse Representation Theory (SDRT) [1]. Les arbres d'annotations que nous cherchons à construire s'appuient sur le versant pragmatique de cette théorie, et permettent de rendre compte de la manière dont les différentes unités d'une conversation s'articulent entre elles.

En outre, les annotations des chercheurs amenés à travailler sur ce projet ne peuvent pas être considérées comme références pour les interprétations : en cherchant des ruptures dans le discours, les "experts" ont potentiellement un comportement biaisé par rapport aux annotations. La subjectivité est inévitable au cours des passations et c'est cela même que nous demandions aux annotateurs : avoir les résultats d'un maximum d'annotateurs naïfs est nécessaire pour avoir une représentation globale du discours et espérer pouvoir approcher la logique interne des patients atteints de troubles du langage et de la pensée. Il était donc nécessaire de mener une campagne pour récolter un maximum de résultats.

2.2 PRISE EN MAIN DE GLOZZ

Glozz[4] est une plateforme permettant d'annoter des textes en les découpant en unités et en reliant ces unités par différentes relations. Les textes à annoter sont importés depuis le format .txt et nous les découpons au préalable en unités, représentant des phrases ou des morceaux de phrases. Le principe est de relier les unités à des unités qui les précèdent par différents types de relations, selon le guide d'annotation conçu par nos prédécesseures et disponible dans leur rapport [2].

[4] http://glozz.free.fr/

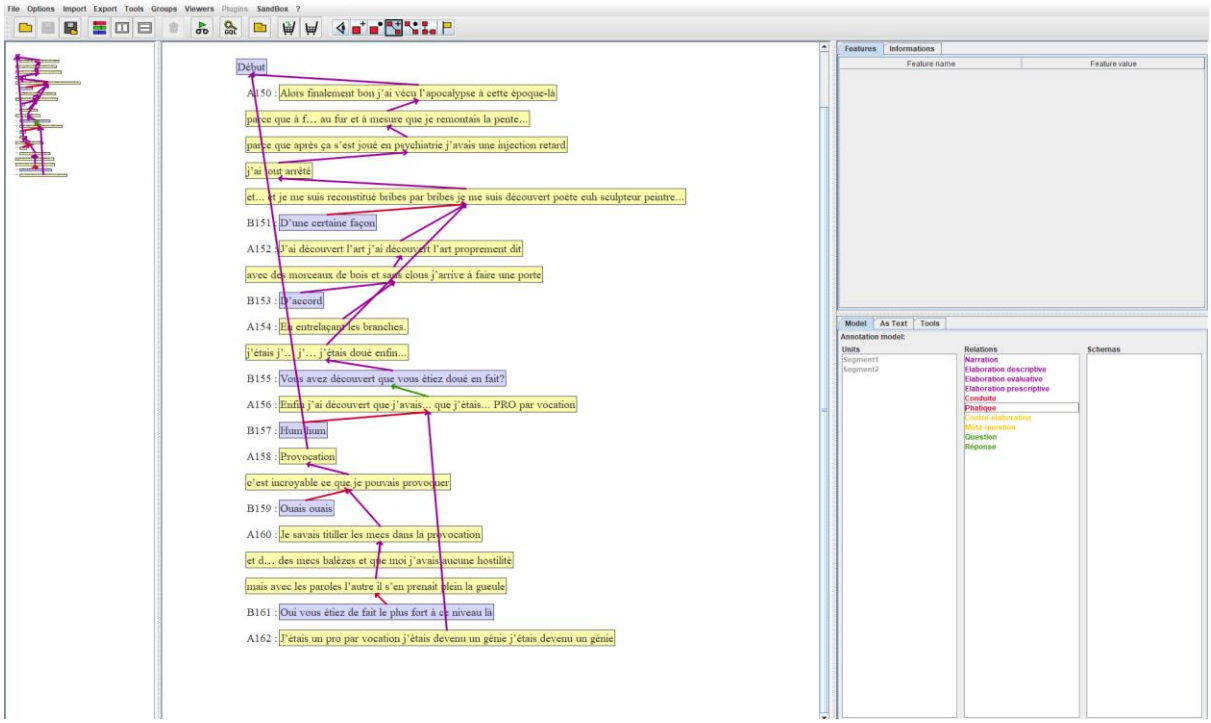


Figure 1 - Interface du logiciel Glozz

Un tutoriel [2] expliquant le paramétrage de Glozz pour définir la liste des types des relations choisies et la pré-annotation des textes en y indiquant les différentes unités avait également été réalisé auparavant.

Pour faciliter l'analyse des données nous avons choisi d'ajouter des caractéristiques aux unités définies dans Glozz : on y ajoute un nom (A1, B2, C3_1, C3_2, etc.) que l'on pourra récupérer directement lors de la lecture des fichiers générés par Glozz par notre script d'analyse. Nous avons également dû changer la manière d'importer les unités pré-annotées avant de faire annoter les relations, afin que les identifiants des unités soient les mêmes pour tous les annotateurs, et que le nom que nous avons ajouté soit présent dans tous les fichiers d'annotations.

En conséquence de ces changements nous avons également pu simplifier l'arborescence des fichiers d'annotations.

La procédure pour créer des unités avec le nom correspondant et notre hiérarchie d'enregistrement des fichiers est disponible en livrables.

2.3 TEXTES ANNOTÉS

Afin d'avoir un point de comparaison lorsque nous voulons détecter une rupture dans un texte, nous avons convenu avec nos tuteurs qu'il serait intéressant d'ajouter à notre panel de textes deux nouveaux textes "témoins" ne comportant pas de rupture discursive mais dont le format de dialogue psy - patient reste assez proche. Nous pourrons ainsi vérifier qu'il y a bien des spécificités dans l'annotation des textes à rupture qu'on ne retrouve pas pour les témoins.

Nos tuteurs nous ont également proposé d'ajouter un nouveau texte à rupture par rapport à la campagne précédente, le texte Sauveur.

En plus du texte "Bac à sable", servant à présenter les consignes et faire prendre en main le logiciel, nous avons donc un total de 6 textes regroupés par paires en fonction de leur type et leur difficulté ou leur longueur :

- Volley et Concours : les textes témoins
- Nord et Florence : les textes à rupture courts
- Sauveur et Provocation : les textes à rupture complexes

Il est important de faire passer les textes de manière aléatoire pour ne pas créer de biais d'annotation par rapport à l'apparition des textes successifs.

Chaque annotateur doit donc annoter un texte (choisi aléatoirement) pour chaque paire, dans un ordre aléatoire. Nous avons donc généré aléatoirement 150 listes de 3 textes avec le script randomizeListe.py, que nous nous sommes ensuite réparties parmi nous trois.

3 CAMPAGNE D'ANNOTATION

3.1 PRÉPARATION / PLANNING / DIFFUSION

Grâce au travail préalable de nos camarades, la préparation de la campagne ne nous a pas demandé une énorme quantité de travail ; nous n'avons pas eu besoin de modifier le protocole des passations ni le guide d'annotation. Seules les améliorations mentionnées dans les parties 2.2 et 2.3 ont été réalisées.

Nous avons planifié des plages horaires de passations en fonction de nos emplois du temps, dans différents lieux (BU, médiathèque, campus …). Pour inviter les gens à s'inscrire, nous avons diffusé en masse des messages sur les réseaux sociaux (essentiellement Facebook) et nous avons créé un évènement sur la plateforme Evento[5]. Malheureusement, nous n'avons pas rencontré le succès que nous espérions par la mise en place de ces dispositifs ; il est probable que le les sujets potentiels n'avaient pas suffisamment de temps (plus de 40min) à accorder pour un sujet aussi complexe, et que par les médias utilisés, les utilisateurs ne sont pas très réceptifs aux annonces.

3.2 CAMPAGNE

Pour la plus grande partie de la campagne, nous avons fait passer des annotations chacun de notre côté. A chaque fois nous n'avons fait passer qu'une ou deux personnes à la fois afin de pouvoir être disponible pour les assister si besoin. Le déroulement d'une annotation est le suivant : on commence par faire lire la consigne et la liste des relations à l'annotateur, en apportant des précisions si besoin, puis on lui fait annoter le texte Bac à sable en restant à côté de lui pour répondre à ses questions. Pour simplifier la compréhension de la tâche, nous lui avons d'abord indiqué comment ajouter les relations puis, une fois que toutes les relations étaient placées, nous lui rappelions qu'il fallait aussi indiquer les thèmes en lui montrant comment faire.

Nous avons pu également nous servir des données recueillies lors de la campagne précédente pour lister les relations qui faisaient débat dans Bac à Sable et les confronter à l'annotation réalisée par la personne : nous lui demandons de justifier, pour chaque relation où plusieurs choix sont possibles, pourquoi il a choisi l'un plutôt que l'autre. Cela permet de s'assurer que les annotateurs ont bien compris la consigne ainsi que tous les types de relations possibles.

Nous insistons aussi sur l'inventaire des choix possibles de l'utilisateur, notamment le fait de pouvoir relier une unité au début par manque de choix ou si l'unité en question n'a pas de référence.

Nous présentons ensuite successivement à l'annotateur les 3 textes qu'il a à annoter, d'après le tirage que nous avons réalisé au préalable. Celui-ci n'est autorisé à nous solliciter que pour passer au texte suivant, pour résoudre un problème technique, ou pour clarifier un point des consignes, mais en aucun cas pour l'aider à décider sur la manière d'annoter.

Avant de lui présenter le texte suivant nous vérifions à chaque fois que toutes les unités sont bien reliées une seule fois et que les thèmes ont été ajoutés.

Une fois les 3 textes annotés, nous faisons remplir à l'annotateur un questionnaire de données personnelles pour pouvoir réaliser une analyse statistique de notre panel d'annotateurs (voir 4.5.1).

[5] https://evento.renater.fr/

3.3 BILAN

Nous avons réussi à recruter 38 annotateurs pour cette campagne 2018, ce qui représente donc, à raison de 3 annotations par personne, un total de 114 annotations différentes (sans compter celles du Bac à sable).

Nous avons eu du mal à toucher des personnes en dehors de nos familles et amis, la diffusion sur les réseaux sociaux n'ayant pas rencontré un grand succès. La méthode la plus efficace a souvent été de prendre contact avec nos connaissances directement et de nous déplacer chez eux pour leur faire passer les annotations, ce qui nous a fait perdre du temps en déplacement et installation.

De plus, les passations duraient en moyenne beaucoup plus de temps que prévu (entre 1h et 1h15 contre 45 minutes estimées à la base. La variance est cependant très élevée, certaines personnes étant beaucoup plus rapides et d'autres ayant passé près de 2h30. Nous avons remarqué que les personnes pour lesquelles la tâche était la plus longue et difficile avaient en fait tendance à surinterpréter les textes plutôt qu'à répondre de manière spontanée.

3.4 AMÉLIORATIONS POSSIBLES

Dans Bac à Sable, plusieurs annotations font apparaître une rupture de la frontière droite : "c'était vraiment super bon !" est rattaché au plat plutôt qu'au restaurant. Il faudrait peut-être discuter avec les personnes concernées pour savoir pourquoi elles le font et leur expliquer en quoi ce rattachement est étrange. Une autre solution serait de trouver un autre texte bac à sable moins ambigu, mais qui permettrait quand même d'avoir de la variabilité dans les annotations.

Beaucoup d'annotateurs se sont plaints de ne pas pouvoir visualiser simplement dans Glozz le type des relations qu'ils venaient d'ajouter, et donc de ne pas pouvoir visualiser leur travail. Glozz ne permet en effet pas d'afficher le nom des relations utilisées, mais une solution simple à mettre en oeuvre serait d'attribuer une couleur différente à chaque relation.

Il serait également intéressant de proposer des relations "neutres" pour que les annotateurs puissent signifier que deux unités sont en relation même s'ils n'en connaissent pas le type, ou si le type ne fait pas partie de ceux proposés (attention toutefois à ce qu'ils n'en abusent pas)

La plupart des annotateurs ne comprennent pas ce qu'on attend des thèmes : ils ont du mal à trouver le bon niveau de détail et certains se servent même plus des thèmes pour analyser le comportement des interlocuteurs ("Inversion des rôles", "Il donne un alibi", …) que pour parler du contenu sémantique de la conversation. Il faudrait sans doute mettre des exemples illustrant des changements de thèmes et les thèmes associés.

Enfin, le guide n'est pas très clair au niveau de l'explication des relations "Question", "Méta-question" et "Réponse" : une réponse ne suit pas forcément une question, elle peut également suivre une méta-question, et une question ne se termine pas forcément par un point d'interrogation. Ces éléments devraient être corrigés.

4 ANALYSE DES RÉSULTATS

4.1 HYPOTHÈSES *A PRIORI*

Suite à la lecture des travaux précédents et aux discussions avec nos tuteurs, nous avons défini des hypothèses assez larges pour nous guider dans nos analyses, correspondant à des pistes qui nous semblaient intéressantes à explorer.

- Certaines relations font plus consensus que d'autres.
- Les passages où les annotations sont le plus dispersées correspondent aux ruptures conversationnelles.
- Il y a des différences dans la manière d'annoter entre les textes témoins et les textes à rupture.
- On peut trouver des catégories d'annotateurs stables d'un texte à l'autre.

Ces hypothèses laissent apparaître 3 grands axes d'analyse : par rapport aux textes, par rapport aux relations, et par rapport aux annotateurs.

4.2 PRÉSENTATION GRAPHIQUE ET OBSERVATIONS

Pour mieux visualiser les données récoltées et nous permettre d'affiner nos hypothèses nous avons dessiné pour chaque texte un graphe "total" présentant les relations indiquées par tous les annotateurs et permettant de visualiser simplement les annotations majoritaires. Pour en simplifier la lecture, nous les avons déclinés en 3 niveaux d'analyse plus ou moins fins, en fusionnant ou pas entre elles certaines relations :

- un graphe représente toutes les relations entre les unités et leur type
- un graphe fusionne certaines relations entre elles selon des catégories définies
- un graphe fusionne tous les types de relations et ne présente donc que l'emplacement des relations

Les catégories que nous avons choisies pour regrouper les différents types de relations sont les suivantes :

- *Narration* : Narration
- *Elaborations* : Elaboration descriptive, Elaboration évaluative, Elaboration prescriptive, Contre-élaboration, Réponse
- *Méta* : Conduite, Phatique, Méta-question
- *Question* : Question

Les graphes présentent également, en rouge, le nombre de personnes ayant signalé un changement de thème au niveau de chaque unité.

En outre, pour plus de lisibilité il est possible de choisir un seuil minimal (en termes de proportion des annotateurs ayant choisi cette relation) pour l'affichage des relations.

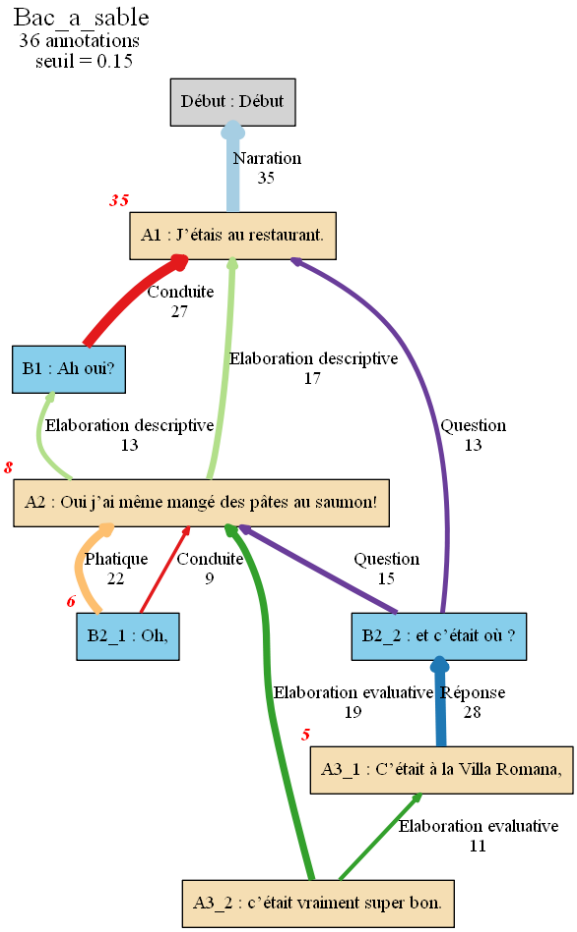


Figure 2- Représentations sans regroupement pour le texte Bac à sable

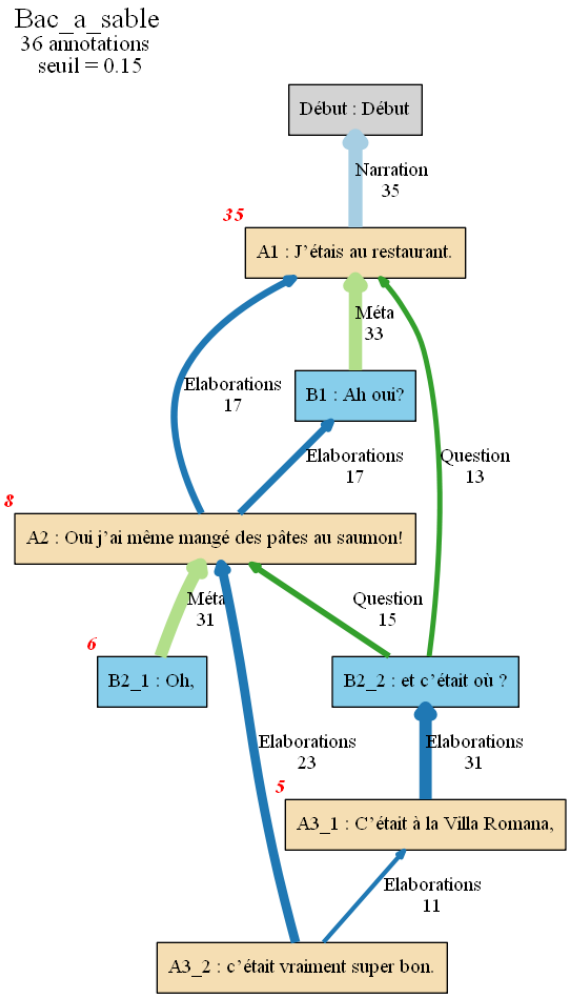


Figure 3 - Représentations avec regroupement par catégories pour le texte Bac à sable

D'après ces différents graphes, il semble que les relations les plus consensuelles soient Question, Réponse et Phatique

Pour tous les textes, il semble que des consensus émergent au niveau de l'emplacement des relations. Les endroits où les rattachements sont les plus dispersés ne semblent pas forcément correspondre aux ruptures conversationnelles que nous avions identifiées, mais plutôt à des unités dont le sens est difficile à saisir (par exemple dans Sauveur : A5_4 : une maison qu'est qu'on a fait qu'on a fait… qu'on a touchée.). On observe d'ailleurs un consensus sur le type de relation utilisé pour relier cette unité (élaboration descriptive) et elle est toujours reliée à une unité proche (on pourrait dire que les différentes annotations sont équivalentes par transitivité).

Dans les textes à ruptures, les annotations semblent être plus consensuelles pour les unités correspondant au psychologue que celles correspondant au patient, dans la mesure où les interventions du psychologue se résument souvent à des phatiques ou des conduites.

A partir des graphes individuels nous avions également observé que même dans le texte Bac à sable certains annotateurs font apparaître une rupture de la frontière droite. Il serait donc intéressant de vérifier que ces ruptures sont quand même plus souvent trouvées dans les textes à ruptures que dans les textes témoin et Bac à sable.

4.3 HYPOTHÈSES À TESTER

A partir de nos idées préalables et de toutes ces observations, et en considérant ce qu'il nous semblait possible de réaliser ou pas avec les moyens et le temps que nous avions, nous avons défini plus précisément une liste d'hypothèses intéressantes à tester.

4.3.1 Hypothèse H1 : Il existe des différences dans la manière d'annoter entre les textes à rupture et les textes témoins

Notre première hypothèse est qu'il existe des différences dans la manière d'annoter entre les textes à ruptures et les textes témoins. En effet comme nous l'avons vu, les textes à rupture, comme ce nom l'indique, ont été choisis parce que les experts y ont détecté des ruptures discursives, comme précisé en 2.1. Ces ruptures sont de deux types : rupture de la frontière droite, et remontées sans complétude de la sous-structure (débrayage conversationnel). Il serait donc intéressant de vérifier si les annotations des non-experts font également apparaître ces deux types de ruptures. Nous pouvons donc décliner les deux sous-hypothèses suivantes :

- H1.1 : Les annotations des textes à ruptures comportent plus souvent des ruptures de la frontière droite que celles des témoins.
- H1.2 : Les annotations des textes à ruptures comportent plus souvent des remontées sans complétude de la sous-structure que celles des textes témoins.

Nous avons également convenu que, les textes à ruptures étant plus difficiles à interpréter et à annoter, il est possible que les ruptures soient manifestées dans les annotations des non-experts par une plus grande dispersion des annotations. En effet, une absence de consensus à ces endroits précis pourrait être un moyen pour les annotateurs de signifier que quelque chose "d'anormal" se passe sans qu'ils soient pour autant capables d'en rendre compte correctement avec le formalisme utilisé. Cette hypothèse (une plus grande dispersion des annotations au niveau des ruptures) est néanmoins difficile à tester, étant donné qu'il est compliqué d'identifier précisément et avec certitude les unités où se trouvent les ruptures et que, comme nous l'avons vu avec nos observations sur les graphes totaux, certains endroits où les annotations sont dispersés correspondent également à des unités qui n'entraînent pas de rupture conversationnelle, mais dont le sens est difficile à identifier. Nous avons donc choisi de tester une hypothèse plus faible, en testant le consensus au niveau des textes plutôt que des unités :

- H1.3 : Les annotations sont plus consensuelles pour les textes témoins que pour les textes à ruptures.

4.3.2 Hypothèse H2 : L'annotation des interventions du psychologue est plus consensuelle que celle des interventions du patient

D'après nos observations dès les premières annotations, incluant même les nôtres lors de la familiarisation avec le logiciel et l'action d'annoter, nous avons estimé qu'il devait exister des différences notables entre l'annotation des unités correspondant aux tours de dialogue du psychologue et celles des tours de parole du sujet schizophrène. Il semble de prime abord normal que le psychologue n'inclut pas de ruptures discursives dans ses interventions, et en effet si celles-ci sont moins nombreuses et plus succinctes, elles semblent aussi se rattacher souvent au groupe de relations conduites/phatiques/questions. Nous émettons donc l'hypothèse que l'annotation des unités du psychologue fera bien plus souvent consensus que celle des unités du patient pour nos annotateurs naïfs.

4.3.3 Hypothèse H3 : Certains types de relation sont plus consensuels que d'autres

En prenant en compte la structure des dialogues entre le psychologue et le patient, ainsi que les types de relations à la disposition des annotateurs, nous estimons que certaines de ces relations seront plus faciles à identifier que d'autres du fait de leur définition et de leurs apparitions dans les textes, en pensant notamment aux questions dont la définition est très claire et le format d'apparition

aussi. Nous émettons donc l'hypothèse plus large que certaines relations sont plus consensuelles que d'autres dans chaque texte.

4.3.4 Hypothèse H4 : On peut regrouper les annotateurs en fonction de la proximité de leurs annotations

Nous avons émis l'hypothèse qu'il est possible de regrouper les annotateurs en fonction de plusieurs caractéristiques de leurs annotations et à partir de ces groupes de pouvoir retrouver des annotations majoritaires qu'on considère comme correctes. Il y a ensuite deux choses à tirer de ces observations :

- H4.1 : L'annotation des experts se trouve dans le plus gros cluster.
- H4.2 : On peut trouver des clusters d'annotateurs stables d'un texte à l'autre.

Si sur un texte donné on a un groupe d'annotations qui fait consensus et qu'on considérera comme la/les bonne(s) annotation(s), alors on souhaite retrouver l'annotation dite experte dans ce cluster majoritaire afin d'affirmer que notre vision de l'annotation experte, donc notre détection de la rupture est bien la bonne annotation.

Ensuite on estime que si un annotateur se trouve dans le cluster majoritaire sur un texte, il a des chances de se retrouver dans le cluster majoritaire sur un autre texte et donc qu'il est un annotateur "fiable", sa bonne annotation n'est pas le fruit du hasard mais bien d'une bonne compréhension des règles d'annotation et du dialogue.

4.4 TECHNOLOGIE ET PROCÉDURES D'ANALYSE

4.4.1 Technologie utilisée

Nous avons effectué nos analyses avec le langage Python et les bibliothèques d'analyse Pandas et Scipy, qui est la solution qui avait été choisie par nos prédecesseures et qui nous a parue pertinente. Ainsi nous avons pu nous inspirer du code qu'elles avaient produit, mais nous n'avons pas travaillé directement à partir de leurs scripts, que nous n'avons pas réussi à faire fonctionner (certains fichiers étaient absents) et qui n'étaient pas très génériques.

Nous avons donc recréé une structure de données plus générique, ainsi qu'un script permettant de lire les fichiers XML générés par Glozz pour les traduire dans cette structure, et écrit un certain nombre de fonctions d'analyse s'appuyant sur cette structure. Nous avons également essayé de documenter au mieux notre code et son utilisation afin qu'il puisse être réutilisé facilement par la suite, même si la campagne change un peu.

Enfin, pour l'exécution de nos scripts d'analyse, nous avons choisi d'utiliser un notebook Jupyter. Cela nous permet d'entrecouper notre code de commentaires sur les résultats, et également de pouvoir faire varier facilement les paramètres de nos fonctions d'analyses pour en observer les résultats. Grâce à cet outil, même des personnes non familières avec la programmation peuvent donc manipuler nos scripts pour explorer les données. Le lecteur désireux d'étudier plus de graphiques que ceux présents dans ce rapport est donc invité à s'y référer.

4.4.2 Stratégie de représentation des annotations

Les annotations produites prennent la forme d'un arbre dont les nœuds sont les différentes unités du texte et les branches sont étiquetées pour représenter le type de relation qui relie deux unités. Cette représentation est utile pour étudier la structure globale des annotations et détecter les ruptures par exemple, mais elle est difficile à utiliser pour comparer les annotations entre elles. Pour nos analyses nous avons donc représenté les arbres d'annotations sous une autre forme.

Chaque unité n'étant reliée qu'une seule fois, on peut considérer qu'une annotation consiste à attribuer à chaque unité une "étiquette", correspondant au type de relation utilisé pour la relier et

à l'unité à laquelle elle est rattachée. Ainsi, une annotation correspond à l'association d'une étiquette pour chaque unité, et il est aisé de comparer deux annotations entre elles en comparant les étiquettes qui ont été choisies pour chaque unité.

Il est en fait possible de choisir différentes façons d'étiqueter les unités, selon qu'on veut comparer seulement le type de relation utilisé par deux annotateurs ou l'unité de rattachement. Pour un maximum de précision sur nos calculs et nos recherches de consensus, nous avons décidé de représenter selon 5 différents niveaux de profondeur les choix des annotateurs pour chaque unité :

- L'unité rattachée
- La catégorie de relation de rattachement (comme définie dans la partie 4.2)
- Le type de la relation de rattachement
- Une combinaison unité rattachée - catégorie de relation de rattachement
- Une combinaison unité rattachée - type de relation de rattachement

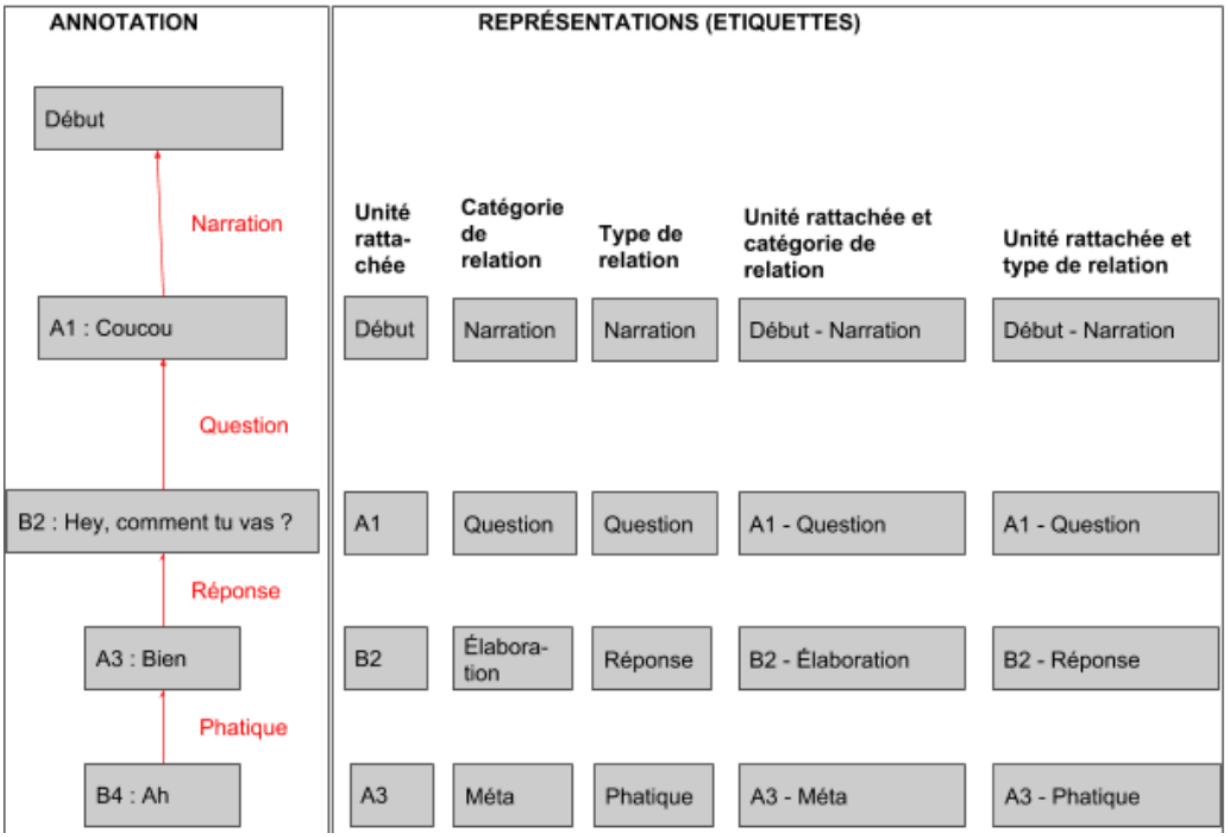


Figure 4 - Exemple de la représentation d'une annotation et des différents types d'étiquetage

Ainsi pour une annotation donnée, on peut construire 5 listes d'étiquettes différentes, mais toutes de même taille néanmoins : le nombre d'unités du texte.

4.4.3 Mesure de consensus

La plupart de nos hypothèses portent sur la notion de consensus sur les annotations, que ce soit au niveau des textes, des unités ou des types de relation. De plus, la notion de convergence est assez floue : il peut s'agir d'une convergence vis-à-vis des types de relation utilisés, ou des unités auxquelles elles se rattachent. En utilisant la représentation introduite au paragraphe précédent (4.4.2), on peut donc essayer de calculer la dispersion des étiquettes choisies par tous les annotateurs pour chaque unité afin d'avoir une idée de la convergence ou non des annotations. De plus, les 5 types d'étiquetages différents permettent de chercher si les consensus se trouvent plutôt au niveau du type des relations ou de l'unité de rattachement (ou les deux).

Pour quantifier la dispersion des annotations au niveau d'une unité, nous avons alors choisi d'utiliser l'entropie de Shannon. Cette mesure permet de calculer le "désordre" dans un ensemble d'objets étiquetés : l'entropie vaut 0 si tous les objets sont du même type et, plus les types sont dispersés, plus l'entropie est élevée.

Elle est calculée comme suit :

$$H(X) = -\sum_{i=1}^{k} P_i \, log_2(P_i)$$

avec *k* le nombre d'étiquettes différentes possible et P_i la proportion d'éléments étiquetés *i*.

Ainsi, les unités annotées de la même façon par tous les annotateurs auront une entropie très faible, celles présentant 2 façons majoritaires d'annoter auront une entropie un peu plus élevée, et celles pour lesquelles les annotations sont très dispersées auront une entropie encore plus élevée.

4.5 RÉSULTATS

4.5.1 Présentation du panel d'annotateurs

Tous les diagrammes se trouvent en Annexe 3.

Nous avons sollicité au total 38 sujets pour cette campagne, principalement parmi nos proches. Le panel est composé à 51,3% de femmes et 46,2% d'hommes, le reste ne souhaitant pas donner l'information.

Nous avons regroupé les annotateurs en catégories d'âges. La catégorie 18-25 ans représente 53,8% du total, la catégorie 25-35 ans représente 7,7%. Le panel ne contient que 2,6% de la catégorie des 35-50 ans. Les 50-60 ans représentent 23,1% des annotateurs et enfin les plus de 60 ans en représentent 12,8%.

Il est intéressant de noter qu'il peut y avoir un biais de représentativité de la population, au vu de la grande part qu'occupe la catégorie des 18-25 ans et la part très faible que représentent les 35-50 ans.

100% de nos annotateurs ont la langue française comme langue maternelle.

Nous avons réussi à obtenir un échantillon diversifié socio-professionnellement, avec moins de la moitié des sujets étant étudiants (46,2%). La seconde catégorie la plus populaire est celle des professions intermédiaires et cadres moyens (20,5%). En troisième viennent les employés et professionnels de service (12,8%), puis les artisans, commerçants, chefs d'entreprises, profession libérales (7,7%). Le reste représente des catégories très minoritaires.

La totalité des participants est diplômée du baccalauréat. 23,7% s'en sont tenus à ce diplôme, 28,9% sont titulaires d'un Bac + 2, c'est à dire IUT, BTS ou DEUG essentiellement. 39,4% sont titulaires d'une licence, et 7,9% sont diplômées d'un Master.

La plus grande part de nos annotateurs vient d'un cursus en informatique ou mathématiques (32,4%). La seconde part la plus conséquente est diplômée de sciences humaines et sociales (29,7%). Les praticiens viennent ensuite, avec 18,9% des annotateurs venant d'une filière professionnelle, et 16,2% provenant des sciences naturelles et techniques. Le reste provient de diverses formations minoritaires.

4.5.2 Recherche de différences entre les textes témoins et les textes à rupture

4.5.2.a Ruptures de la frontière droite

Algorithme de détection de rupture de la frontière droite

Pour pouvoir étudier les ruptures de la frontière droite, il nous a fallu implémenter un algorithme spécifique. Son principe est le suivant : nous parcourons l'arbre d'annotation en utilisant un parcours préfixe, qui consiste à lire un nœud, puis en priorité son fils le plus à gauche, et son nœud frère s'il n'a pas de fils. Dans nos arbres d'annotation, chaque nœud pouvant avoir des fils verticaux ou horizontaux, selon que la relation utilisée correspond à une relation horizontale ou horizontale [4], il convient de traiter en premier ses fils verticaux puis les horizontaux.

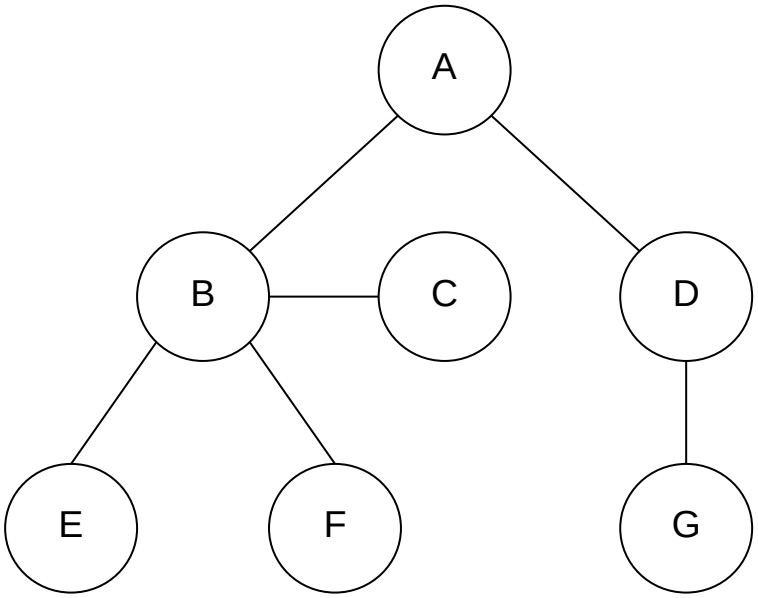


Figure 5 - Exemple d'arbre : l'ordre de parcours préfixe est : A, B, E, F, C, D, G

Nous comparons ensuite l'ordre des unités dans le parcours préfixe avec l'ordre des unités dans le texte. S'il est le même c'est qu'il n'y a pas de rupture de la frontière droite : chaque unité a bien été rattachée à droite de l'arbre. Si les deux ordres sont différents, c'est qu'une unité a été rattachée à gauche d'un de ses parents, rompant ainsi la frontière droite. En parcourant l'arbre en profondeur d'abord on rencontre donc cette unité avant son parent.

Comparaison entre les textes

Pour tester notre hypothèse H1.1 (les annotations comportent plus souvent des ruptures de la frontière droite pour les textes à rupture que les témoins), nous avons compté le nombre d'annotations comportant des ruptures de la frontière droite pour chaque texte.

Tableau 1 - Nombre d'annotations présentant une rupture de la frontière droite pour chaque texte. En rouge les textes à ruptures et en vert les textes témoins

	Annotations comportant une rupture	Annotations totales	Ratio
Bac_a_sable	8	36	0.22
Concours	9	13	0.69
Volley	12	22	0.55
Florence	7	18	0.39
Nord	7	20	0.35
Provocation	10	22	0.45
Sauveur	6	16	0.38

On observe sans surprise que le nombre d'annotations présentant une rupture de la frontière droite pour le texte Bac à sable est très faible, mais les valeurs pour les autres textes semblent assez proches (autour de la moitié des annotations). Nous avons effectué des tests du Chi^2 pour chaque paire de textes afin de vérifier si certaines différences sont significatives.

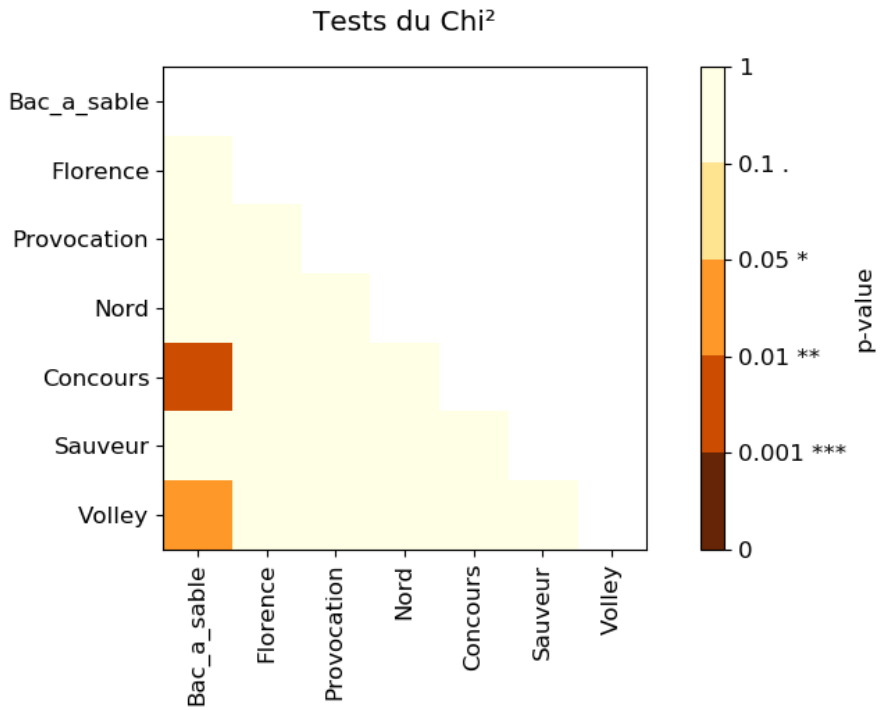


Figure 6 - p-values des tests du Chi^2 deux à deux pour le nombre de ruptures

On n'observe que deux comparaisons significatives au seuil de 5% : entre Concours et Bac à sable, et entre Volley et Bac à sable. Il y aurait donc significativement plus d'annotateurs qui ont fait apparaître des ruptures de la frontière droite dans Volley ou dans Concours que dans Bac à sable.

Ce résultat est très étonnant et contraire à ce que nous attendions, étant donné qu'il s'agit des deux textes témoins qui ne comportaient pas de rupture. On n'observe de plus aucune différence significative que ce soit avec le Bac à sable ou les témoins pour les textes à rupture, on ne peut donc pas dire que les ruptures de la frontière droite y ont été détectées par les annotateurs non experts. Il faut néanmoins nuancer ces résultats, étant donné que le test du Chi^2 est très sensible à la taille des groupes, qui sont ici de taille très modeste (une vingtaine d'annotations par groupe). Il serait donc nécessaire de refaire ces tests après une nouvelle campagne, afin de vérifier si un plus grand nombre d'annotations permettrait de faire apparaître des différences significatives.

4.5.2.b Remontées sans complétude de la sous-structure

Pour tester l'hypothèse H1.2 (les annotations comportent plus souvent des débrayages conversationnels dans les textes à rupture que dans les textes témoins), nous avons procédé de la même façon, en comptant le nombre d'annotations présentant des remontées "illégales" pour chaque texte. Ces remontées se manifestent au niveau des arbres d'annotations par des rattachements à l'unité "Début" d'autres unités que la première du texte. Certains annotateurs ayant tendance à rattacher de nombreuses unités à "Début" et d'autres très peu, nous avons choisi de ne prendre en compte que le nombre d'annotations présentant au moins un tel rattachement, et non pas le nombre de rattachements de ce type par annotation.

Tableau 2 - Nombre d'annotations présentant une remontée pour chaque texte

	Annotations comportant une remontée	Annotations totales	Ratio
Bac_a_sable	1	36	0.02
Concours	1	13	0.07
Volley	1	22	0.05
Nord	6	20	0.30
Provocation	6	22	0.27
Sauveur	8	16	0.50
Florence	13	18	0.72

Cette fois, Florence semble comporter beaucoup plus de remontées sans complétude de la sous-structure que les autres textes. De plus, Bac à sable et les deux textes témoins semblent présenter beaucoup moins de remontée que les autres textes. Nous avons encore une fois effectué des tests du Chi^2 par paire :

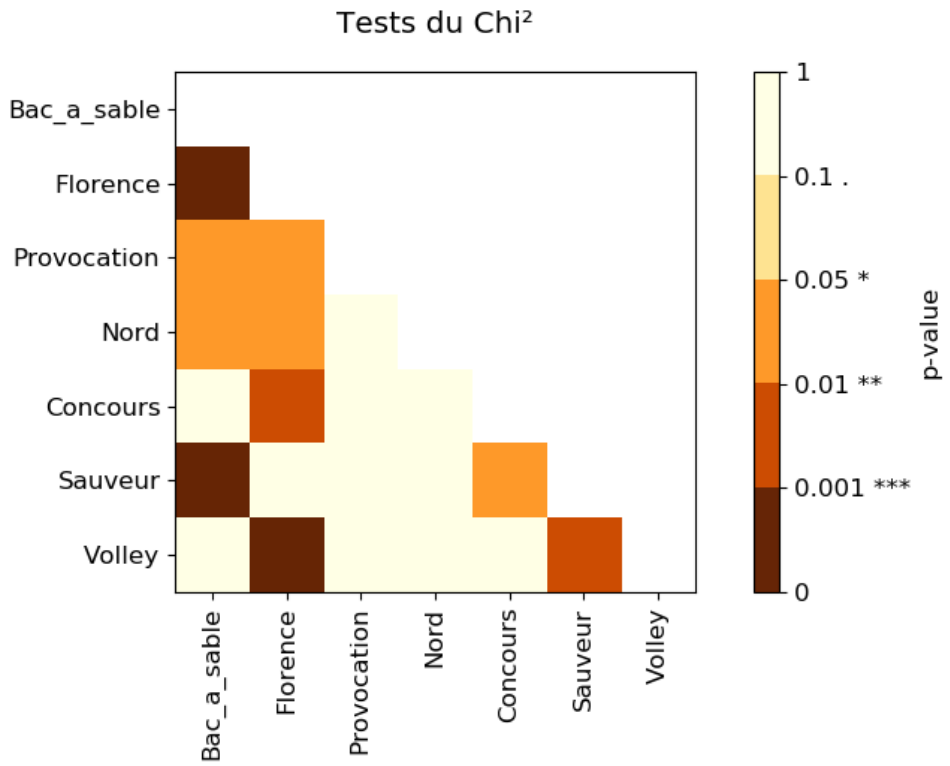


Figure 7 - Significativité du test du chi² sur les remontées sans clôture de la sous-structure

On n'observe pas de différence significative pour le nombre d'annotations comportant des débrayages conversationnels pour le Bac à sable et les 2 textes témoins : dans les trois cas il est proche de zéro. Florence est le texte pour lequel le plus d'annotateurs ont fait apparaître des remontées "illégales", et c'est en effet ce type de rupture qui avait été identifié par les experts sur ce texte. Sauveur présente également significativement plus souvent des rattachements au "Début" que les 3 textes sans rupture. Pour les deux autres textes (Nord et Provocation), la différence avec les textes sans rupture n'est pas significative.

4.5.2.c Consensus des annotations par texte

Notre hypothèse H1.3 est que l'annotation des textes sans rupture est plus consensuelle que celle des textes à rupture. Pour vérifier cela, nous avons calculé l'entropie de Shannon pour chaque unité comme décrit dans le paragraphe 4.4.3, puis nous avons calculé les moyennes d'entropie de chaque texte.

Tableau 3 - Entropies moyennes de chaque texte, selon les 5 critères présentés

	Unité rattachée	Catégorie	Type	Arrivée-catégorie	Arrivée-Type
Bac_a_sable	0.547604	0.391625	0.864300	0.928644	1.338180
Concours	0.642413	0.764375	1.198180	1.218339	1.533880
Volley	0.653012	0.714696	1.197015	1.267928	1.683453
Nord	0.724568	0.716492	1.057102	1.329302	1.639408
Provocation	0.756624	0.781477	1.174379	1.447615	1.774271
Sauveur	0.742390	0.688891	1.098704	1.317406	1.589954
Florence	0.745240	0.821355	1.210562	1.440595	1.722932

On observe une entropie moyenne légèrement plus faible au niveau de l'emplacement des relations pour le texte Bac à sable et les deux textes témoins (Concours et Volley) que pour tous les autres textes, indiquant que ces trois textes sont un peu plus consensuels au niveau de l'emplacement des relations.

Quel que soit le critère, on observe également une entropie beaucoup plus faible pour le texte Bac à sable que pour tous les autres, mais cela s'explique sans doute en grande partie par le fait que ce texte est beaucoup plus court que les autres et qu'il y a donc moins d'annotations différentes possibles.

4.5.2.d Bilan de l'Hypothèse H1

Nous avons essayé de chercher si les différences détectées par les experts entre les textes témoins et les textes à ruptures se retrouvaient au niveau des annotations des annotateurs naïfs. Pour cela nous avons cherché des différences à trois niveaux entre les deux types de textes : au niveau de la présence ou non de ruptures de la frontière droite (H1.1), de la présence ou non de rattachement à l'unité "Début" (H1.2) et au niveau de la convergence globale des annotations par texte (H1.3). Il s'avère que les annotations de non-experts ne font pas apparaître plus de ruptures de la frontière droite pour les textes à rupture que pour les textes témoins, contrairement à ce qui était attendu. Il semble néanmoins que les textes à ruptures se caractérisent par une plus grande divergence des annotations au niveau de l'emplacement des relations (en comparant l'unité rattachée à chaque unité), et par un plus grand nombre de remontées sans clôture de la sous-structure pour certains.

4.5.3 Recherche de consensus sur les interventions du psychologue

Notre hypothèse H2 indique que l'annotation des unités correspondant aux tours de parole du psychologue est plus consensuelle que celle correspondant à ceux du patient. Nous avons donc pour chaque texte dressé la liste de l'entropie des unités correspondant au psychologue et au patient.

Nous avons ensuite essayé de déterminer s'il existait des différences significatives entre les différentes moyennes.

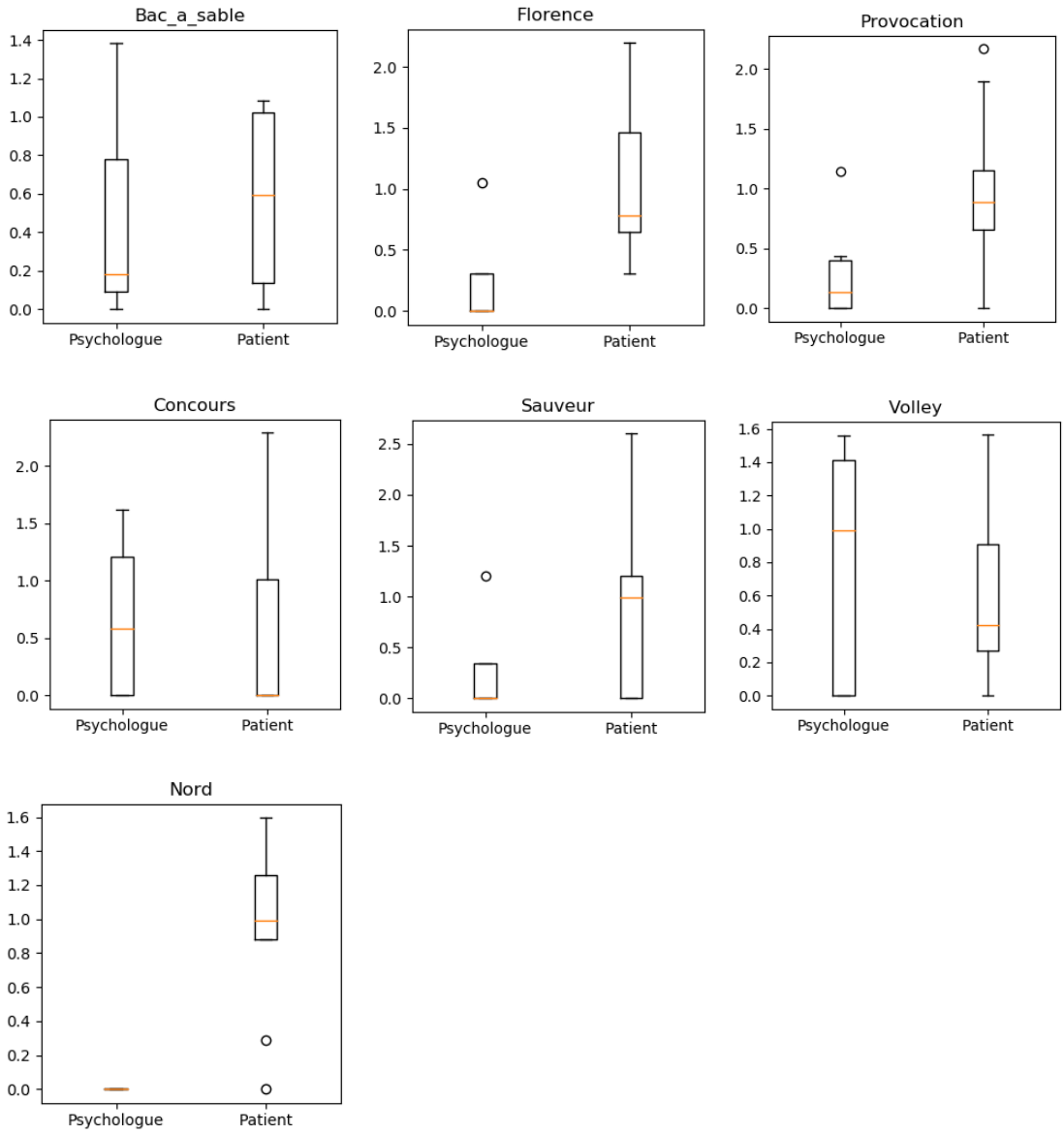


Figure 8 - Entropie par rapport à l'unité rattachée pour les unités du psychologue et du patient pour chaque texte

Il semble que pour tous les textes à ruptures, l'entropie est plus faible pour les unités du psychologue que du patient, si l'on considère l'unité de rattachement de chaque unité.

Nous avons effectué des tests de Wilcoxon-Mann-Whitney pour vérifier si ces différences sont significatives :

Tableau 4 - p-values des tests de Wilcoxon-Mann-Whitney

Bac_a_sable	0.428568
Concours	0.429999
Volley	0.309780
Florence	0.018984
Nord	0.015212
Provocation	0.020644
Sauveur	0.079858

Ces tests concluent à une différence significative pour tous les textes à rupture, sauf Sauveur. Ces différences ne sont néanmoins significatives qu'en calculant l'entropie de manière la plus lâche (c'est-à-dire sur l'emplacement des relations simplement).

Il semble toutefois que notre hypothèse soit validée : les interventions du psychologue sont bien annotées de manières plus consensuelles que celles du patient en ce qui concerne l'unité de rattachement des relations.

4.5.4 Recherche de consensus sur les types de relations

Notre hypothèse H3 est que certains types de relations (i.e. les Questions, les Réponses et les Phatiques) donnent plus souvent lieu à des consensus que les autres. Pour tester cela, nous avons tout d'abord essayé de comparer la proportion de chaque type de relation sur toutes les unités où elle était présente. Pour chaque type de relation, nous avons fait la liste de toutes ses proportions différentes de 0 et nous avons essayé d'en analyser la distribution. Par exemple, si "Narration" correspond à 100% des relations qui partent de A1, 0% de B2, 15% de A3, etc. en retirant les proportions nulles, nous lui assignons la série [1, 0.15, ...]. Les différents textes comportant chacun plus ou moins d'un type ou d'un autre de relations, nous avons choisi de fusionner les résultats de tous les textes.

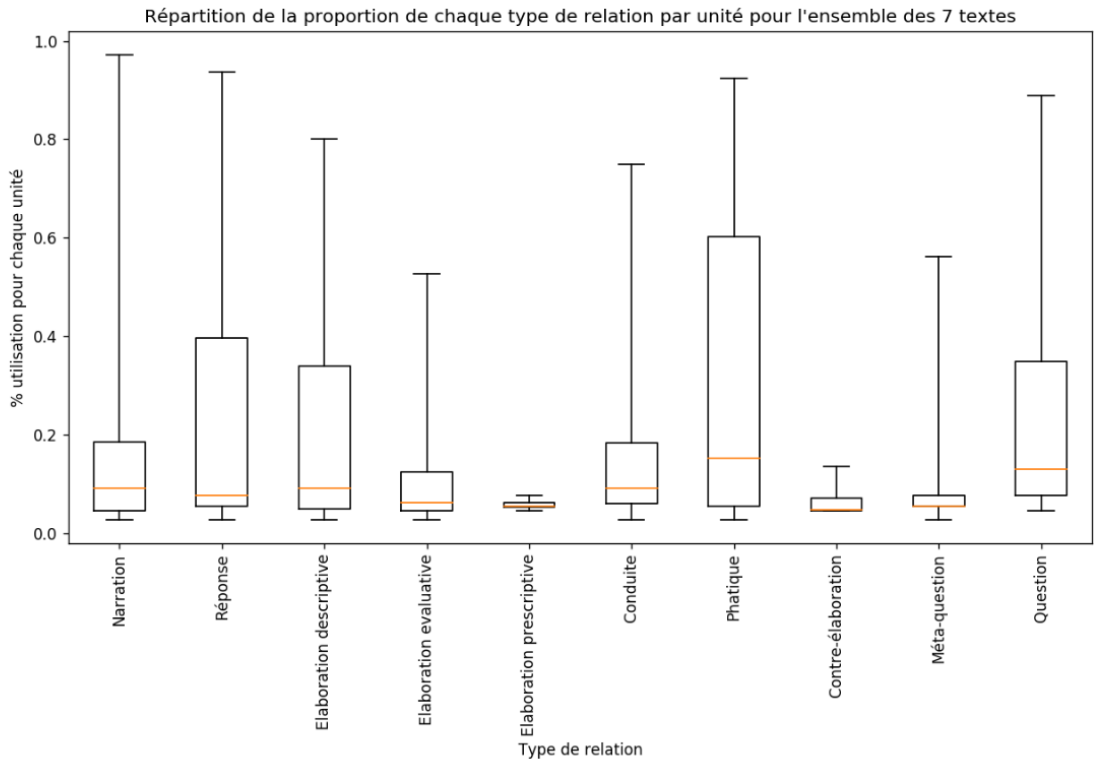


Figure 9 - Répartition de la proportion de chaque type de relation par unité pour l'ensemble des 7 textes

On observe ici que les relations Phatique, Réponse, et également Question et Elaboration, quand elles sont utilisées pour unité, le sont souvent par une plus grande proportion d'annotateurs que les autres types de relation. Par exemple, dans 25% des cas où la relation Phatique est utilisée (le quartile supérieur), elle est choisie par plus de 60% des annotateurs. On observe également que l'élaboration prescriptive n'est presque jamais utilisée, on ne la prendra donc pas en compte dans la suite de l'analyse.

Nous avons effectué des tests sur la médiane afin de vérifier si certains types de relations sont, pour la plupart des unités où ils sont présents, choisis par un plus grand nombre d'annotateurs que d'autres types de relations. Nous avons fait pour cela des tests de Wilcoxon-Mann-Whitney 2 à 2. L'hypothèse nulle est à chaque fois que les distributions des deux groupes sont égales et l'hypothèse alternative que le premier groupe est stochastiquement supérieur au second. Voici les résultats de ces tests (une p-value inférieure à 0.05 indique qu'on peut rejeter l'hypothèse nulle avec moins de 5% de chance de se tromper, et donc accepter l'hypothèse qu'une relation est utilisée dans la plupart des cas conjointement par un plus grand nombre de personnes que l'autre).

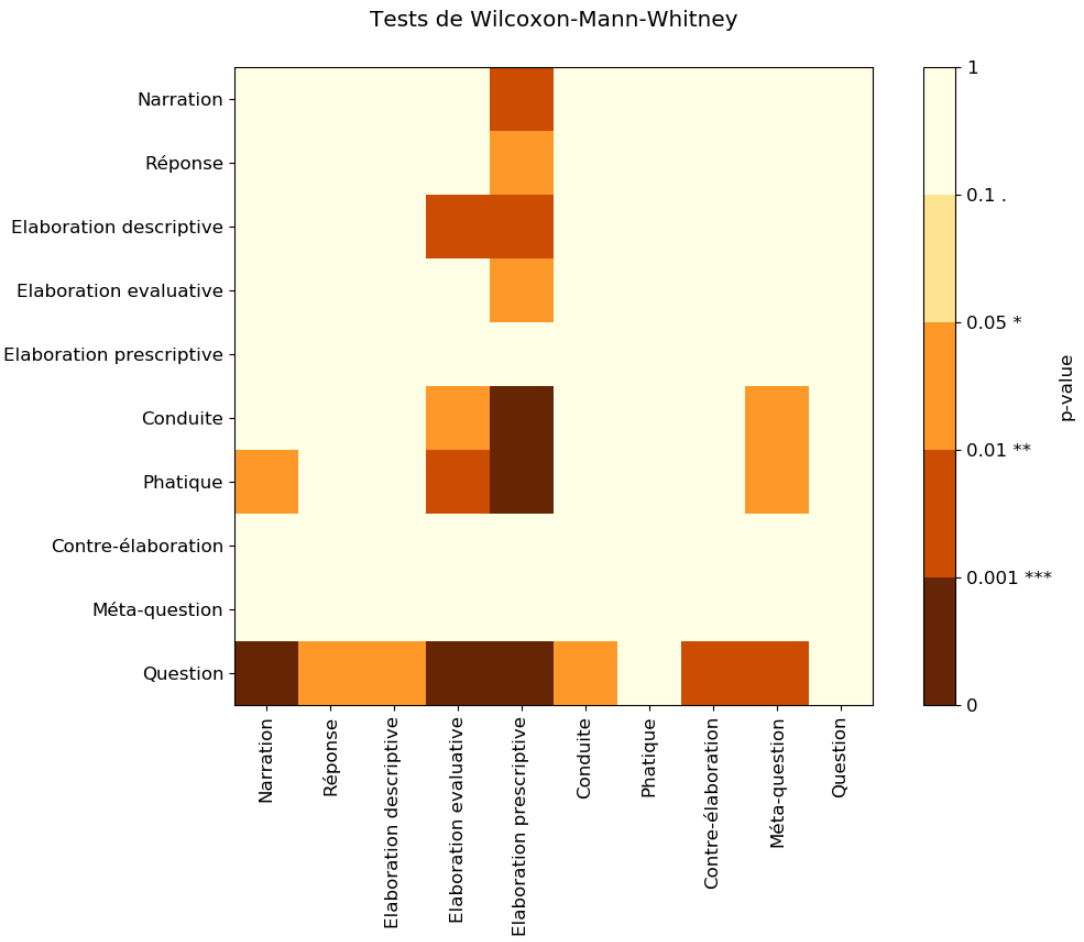


Figure 10 - p-value des tests de Wilcoxon-Mann-Whitney sur la proportion de chaque type de relation par unité

On observe que quand elle est utilisée, la relation Question est choisie par un nombre d'annotateurs significativement plus important que tous les autres types de relation (sauf Phatique). C'est donc sans doute le type de relation le plus consensuel.

Etonnamment, la relation Phatique n'est significativement utilisée conjointement par un nombre plus important d'annotateurs que par rapport aux Narrations, Elaborations descriptives et Méta-questions, alors que c'est la relation pour laquelle la médiane était la plus haute et qui semblait la plus étendue vers les valeurs élevées.

Cependant, ces résultats sont difficiles à interpréter étant donné que les séries contiennent beaucoup de valeurs proches de zéro. En effet, il arrive souvent qu'un annotateur soit seul à choisir un type de relation particulier pour une unité particulière. Il y a donc beaucoup de valeurs proches de 0 mais peu informatives dans la liste des proportions d'utilisation de chaque unité, qui font néanmoins baisser les moyennes et médianes.

4.5.5 Regroupement des annotateurs en fonction de la proximité de leurs annotations

Afin d'estimer des groupes de convergences dans l'annotation des textes par chaque annotateur, il a été décidé de calculer le kappa de Cohen, outil de mesure d'accord inter-juges dont une description est faite dans le texte de Frédéric Santo [5], pour mesurer l'accord deux à deux entre chacune de nos données pour un texte donné et deux annotateurs en utilisant les représentations des annotations en listes décrite dans la partie 4.4.2. Pour créer des sous-groupes d'annotations proches, nous avons besoin de créer des regroupements hiérarchiques d'annotations. Pour cela, nous avons besoin de créer des vecteurs d'annotateurs et de calculer la distance qui sépare leurs annotations

[distance(annotation1, annotation2), distance(annotation1, annotation3), ...)] , générant donc une liste de vecteurs de distances entre chaque paire possible d'annotation.

Pour calculer cette distance, nous avons donc transformé le coefficient de proximité entre deux annotations, calculé par le Kappa de Cohen comme expliqué plus haut, en distance en le soustrayant à 1 : $distance(a1, a2) = 1 - K(a1, a2)$

Avec la liste des vecteurs obtenus, nous pouvons ensuite calculer nos clusters en utilisant les outils de scipy.cluster.hierarchy : linkage pour calculer les clusters en fonction de nos vecteurs de distances et d'un algorithme de regroupement, et ensuite de dendrogram pour les afficher. Nous avons choisi l'algorithme de regroupement "average" car il est un bon compromis entre les différents algorithmes de regroupements.

4.5.5.a Commentaires sur les graphiques :

Il est difficile de donner une valeur minimale générale pour estimer un bon taux d'accord entre deux annotations dû à la variabilité des possibilités de classes d'annotations. En général on donne ces ordres de grandeur au Kappa, mais ces valeurs ne font pas consensus :

Tableau 5 - Ordre de grandeur et interprétation du Kappa

Kappa	Distance correspondante	Interprétation
< 0	>1	Désaccord
0.0 — 0.20	1.0 — 0.80	Accord très faible
0.21 — 0.40	0.80 — 0.60	Accord faible
0.41 — 0.60	0.60 — 0.40	Accord modéré
0.61 — 0.80	0.40 — 0.20	Accord fort
0.81 — 1.00	0.20 — 0.0	Accord presque parfait

On observe généralement et comme nous l'attendions assez logiquement que les clusters avec les distances calculées uniquement sur le choix de l'unité rattachée sont tous beaucoup plus proches que ceux sur les relations de rattachement et eux même plus proche que la combinaison des deux facteurs. Pour chaque texte individuellement nous avons donc pu observer la formation de différents groupes, à distances variables mais globalement une grande fiabilité même dans la nos comparaisons aux niveaux de calculs de distance plus profonds. Celui qui nous intéressera le plus dans les analyses suivantes est la combinaison unité rattachée et catégorie de relation de rattachement, car elle couvre l'entièreté de l'annotation tout en fusionnant les types de relations proches, c'est la plus précise tout en conservant une certaine proximité.

Ces clusters texte par texte ne sont cependant pas très intéressants à regarder seuls et dans l'état, pour répondre à nos hypothèses il va falloir les manipuler.

4.5.5.b H4.1 : L'annotation des experts se trouve dans le plus gros cluster

D'après la définition d'une annotation d'expert décrite dans le rapport du projet tutoré précédent [2], soit l'annotation d'une personne capable d'identifier la rupture dans le discours, et à partir des arbres représentant les annotations expertes pour les textes Provocation, Nord et Florence, nous avons injecté ces représentations dans nos données pour regarder leur position dans les clusters que nous avons formés et ainsi estimer la proximité des annotateurs avec l'annotation experte du texte. Tous les dendrogrammes experts générés sont disponibles en Annexe 4.

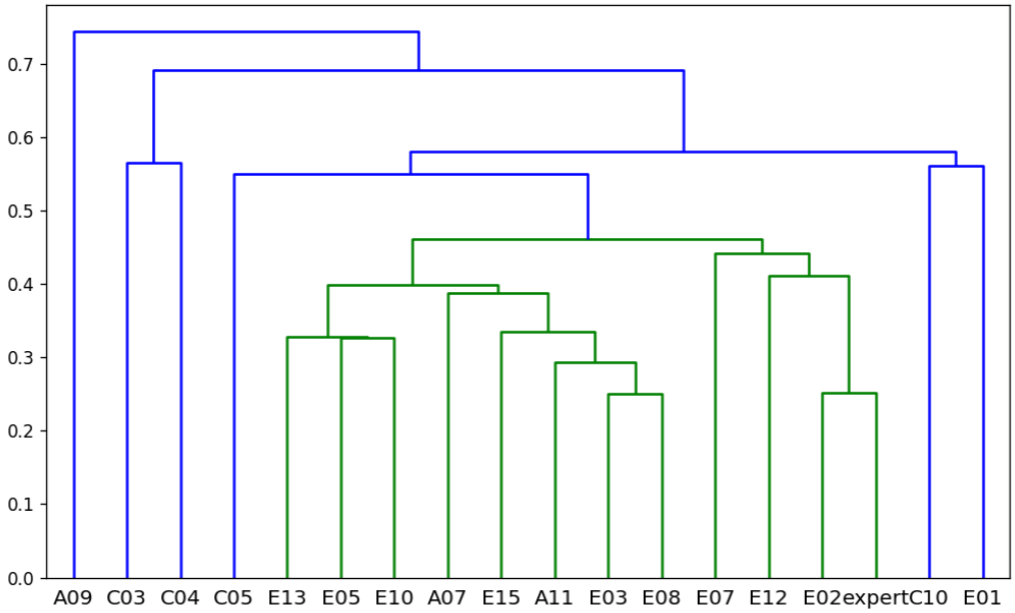


Figure 11 - Dendrogramme Florence avec annotation d'expert

Dans le cas de Florence et Provocation, l'annotation d'expert se trouve dans le groupe majoritaire. Il est important de noter une plus grande disparité dans le très grand groupe de Provocation, l'annotation la plus proche de l'expert est par exemple seulement à 0,4, alors que le texte Florence semble faire un petit peu plus consensus au sein même du groupe. On trouve les annotations de A09, C03 et C04 très éloignées du groupe "expert" de Florence, dans Provocation, on a surtout C02 qui se démarque beaucoup.

Dans le cas de Nord, il y a 5 groupes différents qui se forment, dont le groupe expert constitué de 3 annotateurs autres que l'expert, dont 2 sont très proches, près de 0,2 de distance. Ce petit groupe expert ne rejoint le groupe majoritaire (en bleu) que sous les 0,5 de distance ce qui revient presque à la proximité générale des gros groupes de Florence et Provocation. Ici aussi nous avons C02 qui se démarque beaucoup ainsi que C08 et le groupe A02 et C09.

Nous avons donc pour ces trois textes, la présence de l'annotation experte, donc celle qui a détecté le rupture, dans un groupe majoritaire, ou raisonnablement proche de celui-ci dans le cas de

Nord. Le cas le plus flagrant étant Florence ci-dessus. Ces clusters semblent aussi montrer C02 comme annotateur divergent. Il reste généralement difficile d'estimer un seuil minimal pour estimer le degré de consensus, mais ces groupements proches de l'expert sont bel et bien majoritaires.

4.5.5.c H4.2 : On peut trouver des clusters d'annotateurs stables d'un texte à l'autre

Du fait de la grande variabilité des textes annotés il a été difficile de réunir des données semblables à comparer entre elles sur l'échelle globale de la campagne. Nous avons fait les choix de pouvoir sélectionner les textes à comparer un par un pour simuler le choix des textes lors d'une passation et d'afficher le cluster correspondant au cumul des distances entre chaque annotation de chaque texte.

Le but est de voir si avec ces clusters sur plusieurs textes il y a une consistance dans les groupes formés de textes en texte, notre hypothèse étant que les bons annotateurs restent bons d'un texte à l'autre, les clusters devraient être les mêmes à l'ajout des calculs de distances d'un nouveau texte au cluster. Ne pouvant comparer entre elles que les annotations d'un même texte, plus le nombre de textes sélectionnés conjointement augmente plus le nombre d'annotateurs dans le cluster diminue et moins il est pertinent.

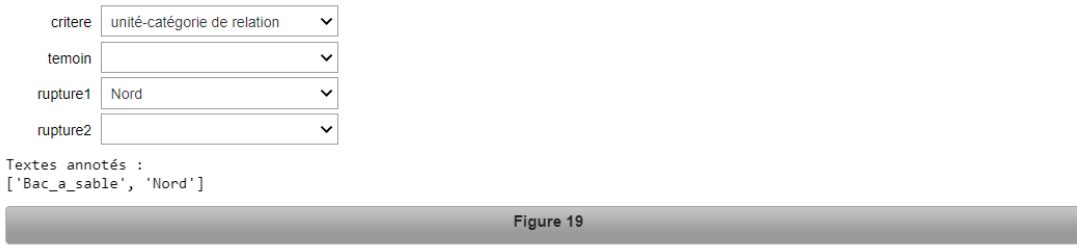


Figure 19

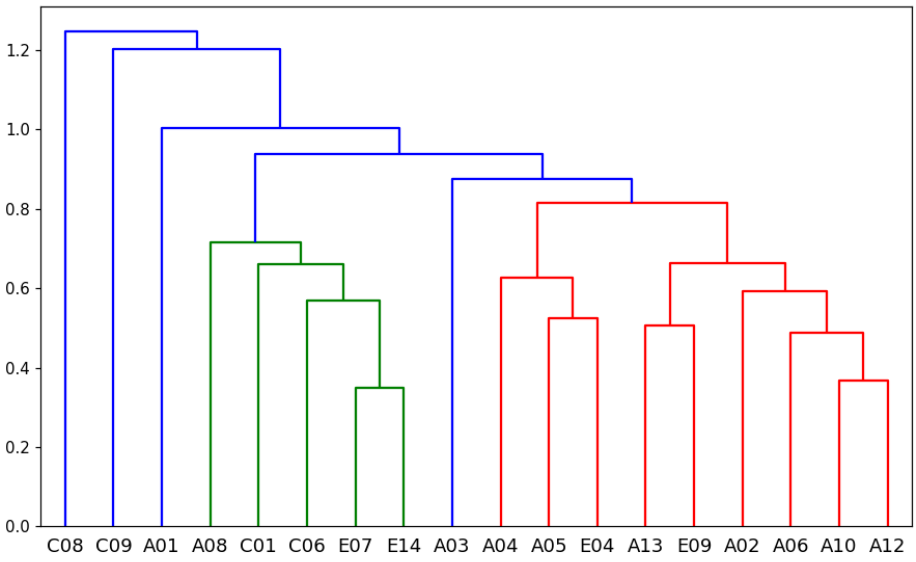


Figure 12 - Dendrogramme des textes Bac à Sable et Nord

critere unité-catégorie de relation
temoin
rupture1 Nord
rupture2 Provocation

```
Textes annotés :
['Bac_a_sable', 'Nord', 'Provocation']
```

Figure 20

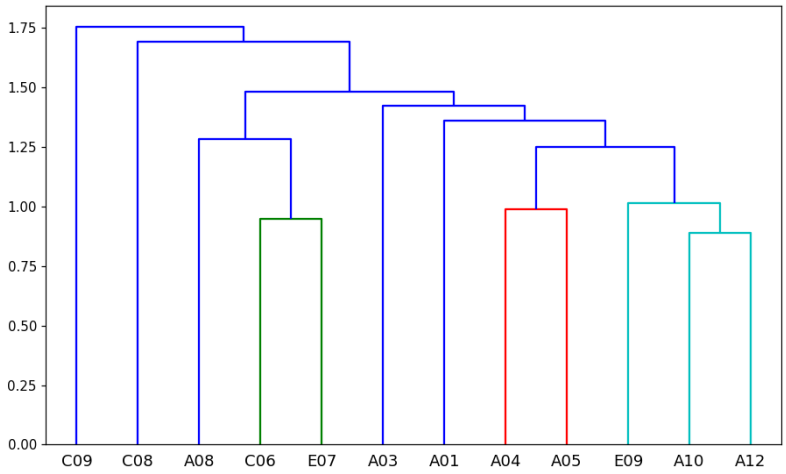


Figure 13 - Dendrogramme des textes Bac à Sable, Nord et Provocation

Analysons par exemple les différents clusters formés pour les textes Bac à Sable et Nord puis leurs variations à l'ajout du texte Provocation.

C02, C09 et A01 sont à part et nous avons deux gros clusters. A l'ajout de Provocation, C02 et C09 sont toujours des annotateurs à part, et nous n'avons plus que 3 petits groupes dont deux se rejoignent assez tôt mais cependant toujours relativement distincts. En identifiant les annotateurs, on remarque que le groupe vert est ce qu'il reste du précédent groupe vert alors que le groupe rouge a été divisé en deux petits groupes rouges et bleus. Il reste donc une certaine cohérence dans la formation de ces groupes.

Cependant notre échantillon est trop petit pour que ses observations soient vraiment concluantes sur toute la campagne et surtout pour l'automatiser, au vu de la variabilité dans la passation des textes.

4.7 BILAN

Nous n'avons pas réussi à trouver de différence significative au niveau de la détection des ruptures de la frontière droite entre les textes témoins et les textes à rupture (section 4.5.2), mais nous avons trouvé des différences en termes de débrayages conversationnels entre certains textes à rupture (Florence et Sauveur) et les textes sans rupture. Quoi qu'il en soit, la principale différence d'annotation entre les deux types de textes semble se manifester par un plus grand consensus sur l'emplacement des relations pour les textes témoins et bac à sable que pour les textes à ruptures, ce qui confirme notre hypothèse H1.3.

Plus précisément, nous avons montré (section 4.5.3) en validant l'hypothèse H2 que les divergences dans les textes à ruptures se trouvent au niveau des interventions des patients, et non pas des psychologues. Il semble donc que les désaccords entre annotateurs soient un moyen de mettre en évidence les anomalies du discours. Dans cette optique, il serait intéressant par la suite de vérifier si ces désaccords ont lieu précisément sur les unités où les experts avaient détecté des ruptures, ou sur n'importe quelle intervention du patient.

Nous avons en outre montré dans la partie 4.5.4 que les relations les plus consensuelles sont celles de type Question et Phatique, ce qui est cohérent avec le résultat précédent étant donné que ces relations sont majoritairement utilisées par le psychologue. Cependant ces résultats étaient difficiles à interpréter et il serait nécessaire d'effectuer d'autres tests et de collecter plus de données pour les consolider.

Pour ce qui est des clusters d'annotateurs, nous retrouvons dans 2 cas sur 3 notre expert dans le cluster majoritaire, et dans le troisième cas il reste dans un groupe suffisamment proche, et si notre technique de randomisation des textes passés rend difficile la généralisation des clusters sur toute la campagne, quelques observations individuelles semblent indiquer une consistance dans la qualité d'annotation d'un texte à l'autre pour un même annotateur. La création de ces regroupements semble donc indiquer la présence majoritaire d'annotateurs ayant une vision des textes Provocation, Nord et Florence proche d'un expert, et une consistance entre les textes que cette faculté de détection de la rupture n'est pas liée à un texte précis. Cependant, cette dernière affirmation nécessite d'effectuer des tests automatisés sur des échantillons bien plus larges.

A l'avenir il serait intéressant d'effectuer d'autres campagnes afin de recueillir plus de données et d'essayer d'obtenir des résultats plus fiables. Le code que nous avons produit est normalement assez générique pour permettre de fusionner facilement les données de plusieurs campagnes.

Enfin, notre algorithme de détection des ruptures de la frontière droite ne permet pas de détecter l'emplacement des ruptures dans le texte, mais seulement leur présence ou non. Une priorité pour les analyses futures nous semble donc être de chercher à comparer le lieu des ruptures entre les annotateurs (rupture de la frontière droite comme débrayage conversationnel), ainsi que de s'intéresser à l'emplacement des changements de thèmes et aux éventuels retours à un thème déjà évoqué

5 CONCLUSION

Notre objectif lors de ce projet tutoré était de mener une campagne d'annotation, facilitée par la mise en place des outils et des tutoriels réalisés par nos camarades travaillant sur le même projet de recherche pendant les années précédentes.

Une part importante de notre mission résidait en outre dans le fait d'analyser en profondeur les données exploitées lors de cette campagne, qui n'ont pas été totalement interprétées lors des derniers projets.

Grâce aux outils dont nous bénéficions, il a été relativement simple d'effectuer la campagne d'annotation. La partie la plus chronophage de cette dernière comportait la mise en place d'une campagne de communication, d'appel à participants. Nous estimons qu'un résultat de 38 annotateurs est un bon score, sachant que tout le monde a joué le jeu et participé jusqu'au bout des passations, ce qui n'est pas une chose *a priori* simple puisque les tests ont une durée moyenne d'une cinquantaine de minutes par annotateur. Toutefois, le nombre d'annotations récoltées aurait peut-être pu être amélioré grâce à une campagne d'annonces plus efficace.

Malgré quelques déchets dans les annotations (dûs à des erreurs de saisie ou autres), la quasi-totalité des données était exploitable.

Ce projet a nécessité de nombreuses connaissances et compétences acquises au fur et à mesure de notre cursus universitaire, alliant programmation, analyse de données, pédagogie et organisation. Nous avons notamment appris de nombreuses choses dans le domaine de la linguistique.

Cet exercice nous a permis de retrouver des connaissances enfouies depuis quelques années de cours puisqu'il nécessitait un grand éventail de compétences. Nous avons dû faire face à quelques difficultés notamment dans la reprise du code des travaux précédant notre arrivée dans le projet, pour l'analyse des données. Par manque de communication, nous avons dû reprendre certaines parties depuis le début, ce qui peut utiliser un temps précieux.

Toutefois, et grâce à ce léger incident, nous avons mis l'accent sur la clarté de notre code pour que nos potentiels successeurs puissent poursuivre aisément nos travaux. Nous espérons également que nos remarques recueillies suite à la campagne d'annotation puissent éventuellement servir à l'amélioration du déroulement de celle-ci à l'avenir.

6 BIBLIOGRAPHIE

[1] M. Amblard, M. Musiol, and M. Rebuschi. Une analyse basée sur la S-DRT pour la modélisation de dialogues pathologiques. In Traitement Automatique des Langues Naturelles - TALN 2011, page 6, Montpellier, France, June 2011.

[2] L. Huber et E.Laurier. Rapport de projet tutoré, Trouble du langage et de la pensée : Campagne d'annotation - 2017.

[3] M. Musiol et M. Rebuschi. La rationalité de l'incohérence en conversation schizophrène (Analyse pragmatique conversationnelle et sémantique formelle). In ScienceDirect - 137-169, 2007.

[4] M. Rebuschi, M. Amblard, M. Musiol. Schizophrénie, logicité et perspective en première personne. In L'évolution Psychiatrique, juin 2011.

[5] Frédéric Santos, Le kappa de Cohen : un outil de mesure de l'accord inter-juges sur des caractères qualitatifs, 4 avril 2018.

7 ANNEXES

Annexe 1. CALENDRIER / DIAGRAMME DE GANTT

Tableau 6 - Diagramme de Gantt

WBS	Task description	Start date	Finish date	Progress
1	Travail bibliographique	13/11/2017	15/02/2018	100%
2	Prise en main de GLOZZ	15/02/2018	28/02/2018	100%
3	**Analyse de données**	**01/03/2018**	**29/04/2018**	**100%**
3.1	**Analyse large des données précédentes**	**01/03/2018**	**31/03/2018**	**100%**
3.1.1	Matrice de contingence	01/03/2018	31/03/2018	100%
3.1.2	Créer une structure de données	15/03/2018	31/03/2018	100%
3.1.3	Points de convergence	01/03/2018	31/03/2018	100%
3.1.4	Clusters utilisateurs	19/03/2018	31/03/2018	100%
3.2	**Analyse ruptures discursives**	**26/03/2018**	**29/04/2018**	**100%**
3.2.1	Algorithme frontière droite	26/03/2018	29/04/2018	100%
3.2.2	Algorithme 2 arbres	26/03/2018	29/04/2018	100%
3.2.3	lien avec les changements de thème	26/03/2018	29/04/2018	100%
4	**Organisation Campagne**	**01/03/2018**	**08/04/2018**	**100%**
4.1	Mise au point du formulaire de retour	01/03/2018	07/03/2018	100%
4.2	Mise au point du formulaire de recrutement	21/03/2018	30/03/2018	100%
4.3	Location d'ordinateurs à l'UFR	21/03/2018	28/03/2018	100%
4.4	Contact des médiathèques etc	21/03/2018	28/03/2018	100%
4.5	Elaboration du planning	21/03/2018	28/03/2018	100%
4.6	Diffusion du formulaire	26/03/2018	08/04/2018	100%
4.7	Contacts et planification des créneaux	26/03/2018	08/04/2018	100%
5	**Campagne**	**09/04/2018**	**06/05/2018**	**100%**
5.1	Passations à Nancy	09/04/2018	27/04/2018	100%
5.2	Passations élargies en campagne	27/04/2018	06/05/2018	100%
6	**Analyse des résultats**	07/05/2018	23/05/2018	100%
7	**Rendu**	24/05/2018	31/05/2018	100%

Annexe 2. LISTE DES TEXTES D'ANNOTATION

Début

A1: J'étais au restaurant.

B1: Ah oui?

A2: Oui j'ai même mangé des pâtes au saumon!

B2: Oh,

et c'était où ?

A3: C'était à la Villa Romana,

c'était vraiment super bon.

Figure 14 - Texte Bac à Sable

Début

B20 : c'était quand même assez stressant euh la / la prépa...

A21 : Mmh mmh.

B23 : donc euh... donc du coup ouais euh et bon pour euh... en ce qui concerne les études

donc du coup après j'ai / j'ai arrêté le / le / le / le / l'école d'ingénieur enfin la pépa...

je suis revenue à Ville1…

A24 : Mmh mmh.

B25 : j'ai fait euh une / une / une fac de / de maths... je suis allé en fac de maths.

A26 : Ouais.

B27 : Euhh et là pareil je / j'étais pas j'avais un / un appart en colocation euh avec deux amis...

euh et / et donc euh j'ai / j'ai échoué en fait

euh... j'ai pas tout échoué j'ai validé certaines matières

mais euh... j'étais pas en forme...

Figure 15- Texte Nord

Début

A1 : Et euh donc là vous avez passé le concours c'est…

B2 : Ouais j'attends les résultats.

A3 : C'est quand les résultats?

B4 : Juin.

A5 : Juin?

B6 : Mmh.

A7 : Parce que je sais qu'en début d'année y en a qui attendent encore les résultats mais euh y a deux sessions de d'exams ou euh.

B8 : Euhh éduc spé les... en début d'année c'est la passation de l'écrit.

On passe l'écrit au mois de janvier.

A9 : Ouais.

B10 : Euh si on est reçu parce que cette année ils étaient vachement en retard.

A11 : D'accord.

B12 : Si on est reçu à l'écrit on passe un oral.

A13 : Mmh mmh.

B14 : Et si on est reçu à l'oral on rentre dans l'école.

A15 : D'accord.

Et y a beaucoup de places?

B16 : 90.

A17 : D'accord.

Vous étiez beaucoup à…

B18 : Ben cette année on était pas beaucoup.

Figure 16 - Texte Concours

Début

B1 : J'aimerais savoir ce que font les personnes qui sont à l'hôpital

ce que vous faites la journée par exemple…

A2 : Je suis très amoureuse de Florence M.

B3 : De Florence M.

A4 : Oui superbe la…

comment elle s'appelle Florence R.

elle a tué quand même plus de un million de de personnes

B5 : Qui ça ?

A6 : Florence R.

B7 : C'est qui cette dame là ?

A8 : Elle était psychiatre 40 rue de N.

j'y allais une fois par semaine ou deux fois tous les quinze jours

elle aurait pu me tuer mais enfin...

Figure 17 - Texte Florence

Début

A150 : Alors finalement bon j'ai vécu l'apocalypse à cette époque-là

parce que à f… au fur et à mesure que je remontais la pente…

parce que après ça s'est joué en psychiatrie j'avais une injection retard

j'ai tout arrêté

et... et je me suis reconstitué bribes par bribes je me suis découvert poète euh sculpteur peintre…

B151 : D'une certaine façon

A152 : J'ai découvert l'art j'ai découvert l'art proprement dit

avec des morceaux de bois et sans clous j'arrive à faire une porte

B153 : D'accord

A154 : En entrelaçant les branches.

j'étais j'… j'… j'étais doué enfin…

B155 : Vous avez découvert que vous étiez doué en fait?

A156 : Enfin j'ai découvert que j'avais... que j'étais... PRO par vocation

B157 : Hum hum

A158 : Provocation

c'est incroyable ce que je pouvais provoquer

B159 : Ouais ouais

A160 : Je savais titiller les mecs dans la provocation

et d… des mecs balèzes et que moi j'avais aucune hostilité

mais avec les paroles l'autre il s'en prenait plein la gueule

B161 : Oui vous étiez de fait le plus fort à ce niveau là

A162 : J'étais un pro par vocation j'étais devenu un génie j'étais devenu un génie

Figure 18 - Texte Provocation

Début

A1: enfin moi où j'ai quand même souffert c'est quand j'ai eu mon traumatisme crânien

B2: C'est dû à un accident ?

A3: Oui enfin on m'avait pratiquement culbuté.

C'est quand même bien un... ç'en est un qui m'a balancé... qu'était devant et puis moi derrière et qui m'a…

mais enfin je m'en fous parce que... enfin j'étais chargé

J'avais quand même 5 litres de vin de... de pineau 5 litres de bière plus 1 ou 2.

B4: Que vous aviez bu ?

A5: On allait dans la ferme à S. (lieu)

une ferme abandonnée.

Qu'habite qu'est... qu'appartient à Henry euh…

une maison qu'est qu'on a fait qu'on a fait... qu'on a touchée.

Moi j'aimais j'aime bien Franck. Franck L (nom).

il m'a sauvé... avec son frère.

B6: Il vous a sauvé comment ça ?

A7: Comment ?

B8: Comment ça il vous a sauvé ?

A9: Ah mais euh... qui euh... ben il m'a sauvé euh parce que j'étais avec lui parce qu'il... buvait quoi.

Il voulait me taper dessus et en plus son frangin il m'a enlevé.

Ben ce qui m'aide là c'est quelqu'un de bien c'est Damien.

Ben heureusement qu'il m'a fait ça parce que... il il faisait comme ça il se faisait disparaître et je peux le faire moi disparaître

B10: Donc vous pouviez disparaître et réapparaître ?

A11: Oui.

Figure 19 - Texte Sauveur

Début

A1 : Je suis ici depuis deux ans et j'en ai fait une dizaine d'années à B. près de M.

B2 : D'accord, ok et ça va vous avez un bon niveau?

A3 : Ouais.

Bon le problème c'est que je suis un peu petit donc euh pour le volley c'est pas pratique.

B4 : Ouais oh…

A5 : Donc je peux faire passeur ou libéro je sais pas si vous connaissez un peu le volley?

Ou euh …

B6 : Alors …

A7 : Pas du tout?

B8 : J'ai regardé je suis tombée sur une émission euh sportive

parce que maintenant ils retransmettent pas mal de matchs…

A9 : Hum hum.

B10 : Euh l'autre soir sur euh je sais pas l'équipe ou je sais plus quelle chaîne …

Et c'était justement la finale euh un match féminin. De volley féminin …

A11 : Hum hum.

B12 : Et je ne comprenais pas pourquoi une nana n'avait pas la même couleur de maillot que les autres.

A13 : C'est justement les libéros en fait

c'est quelqu'un qui fait que les postes arrières …

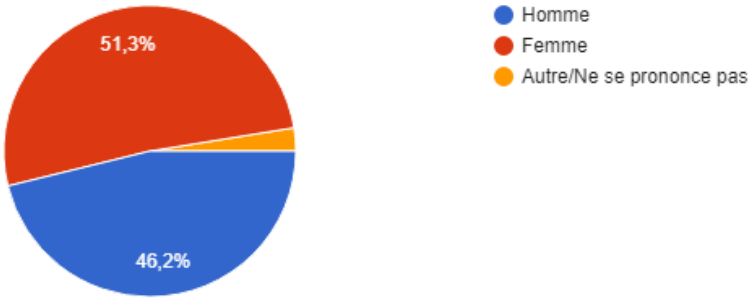
B14 : D'accord.

Figure 20 - Texte Volley

Annexe 3. PANEL D'ANNOTATEURS

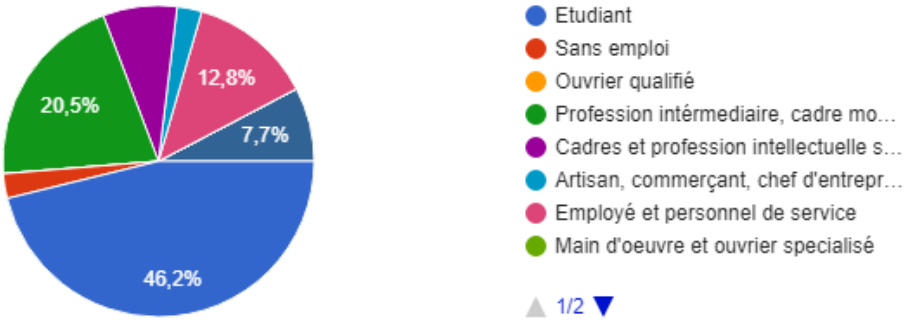
Sexe

39 réponses



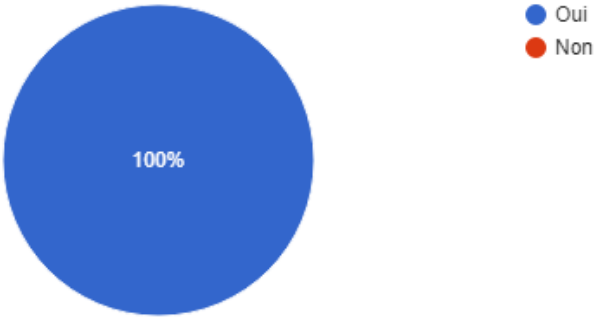
Catégorie Socio-Professionnelle

39 réponses



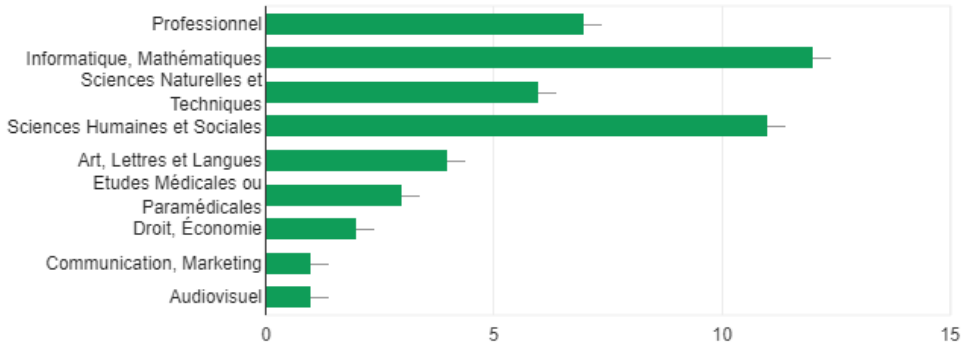
Le français est-il votre langue maternelle ?

39 réponses



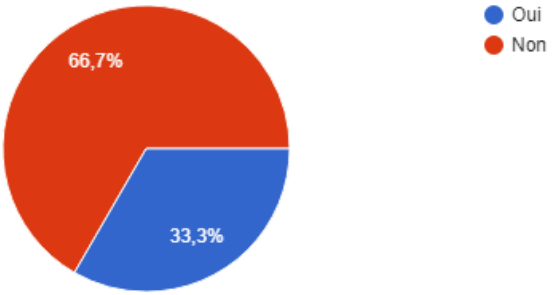
Quel est (ou a été) votre domaine d'étude

37 réponses



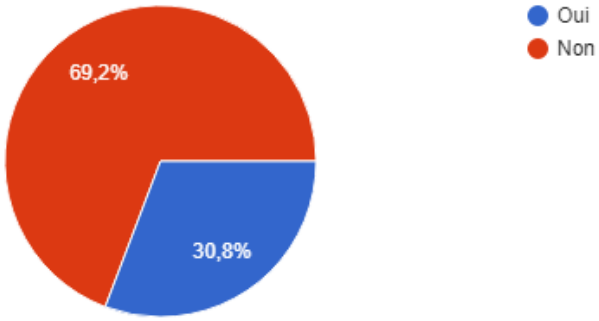
Êtes vous familiers avec les sciences du langage ?

39 réponses



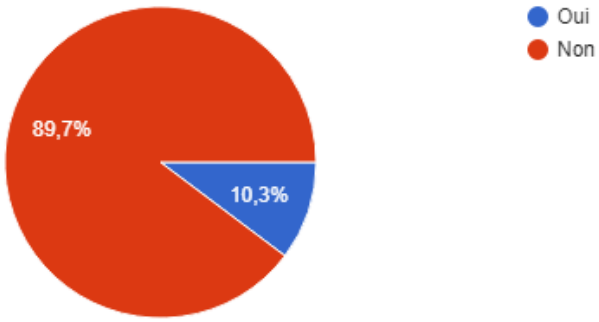
Avez vous déjà participé à des expériences scientifiques ?

39 réponses



Avez vous déjà réalisé des annotations similaires ?

39 réponses



Annexe 4. CLUSTERS AVEC EXPERTS

textname Provocation

critere unité-catégorie de relation

Figure 9

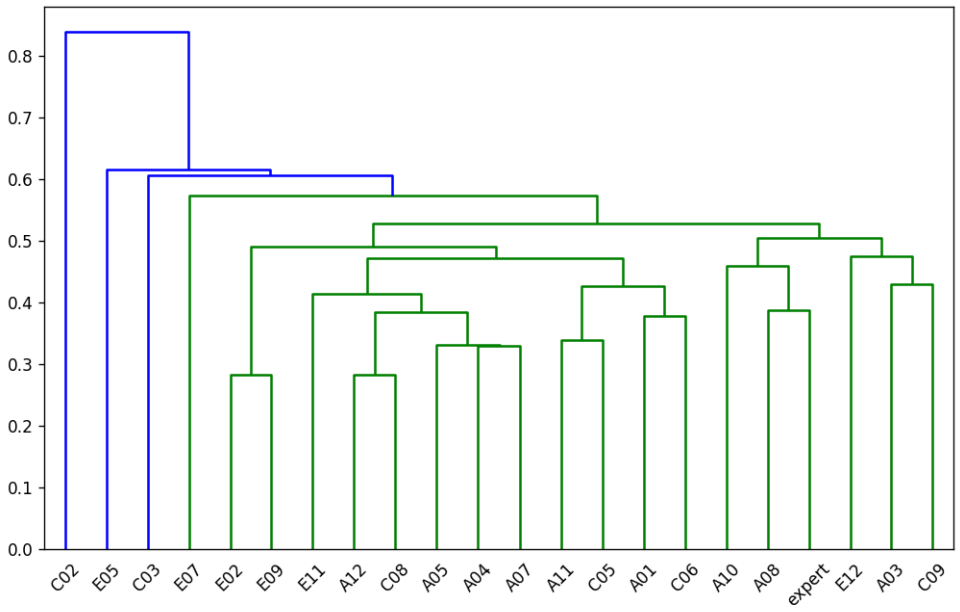


Figure 21 - dendrogramme Provocation avec annotation d'expert

textname Florence

critere unité-catégorie de relation

Figure 8

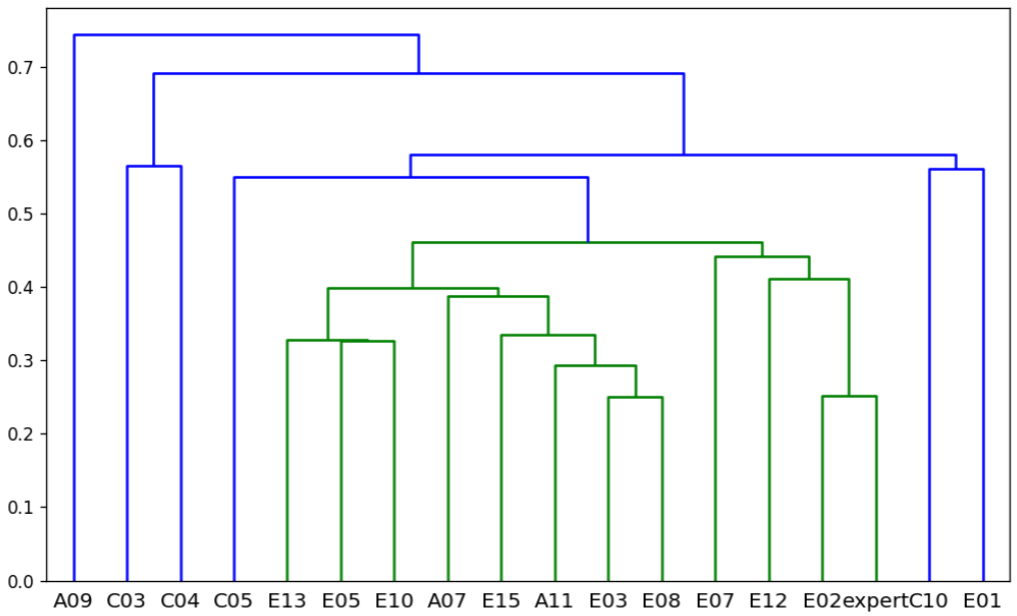


Figure 22 - Dendrogramme Florence avec annotation d'expert

textname Nord

critere unité-catégorie de relation

Figure 10

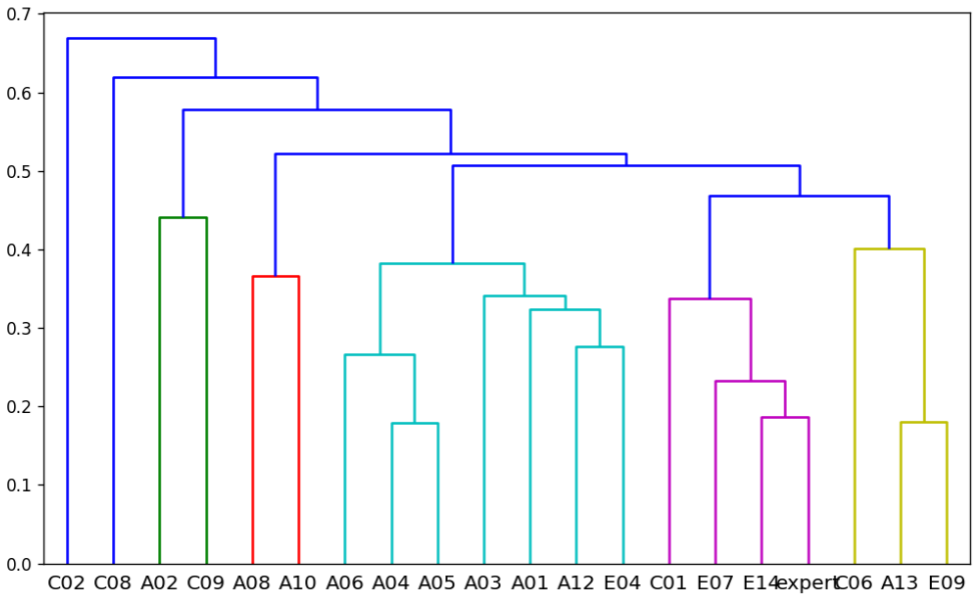


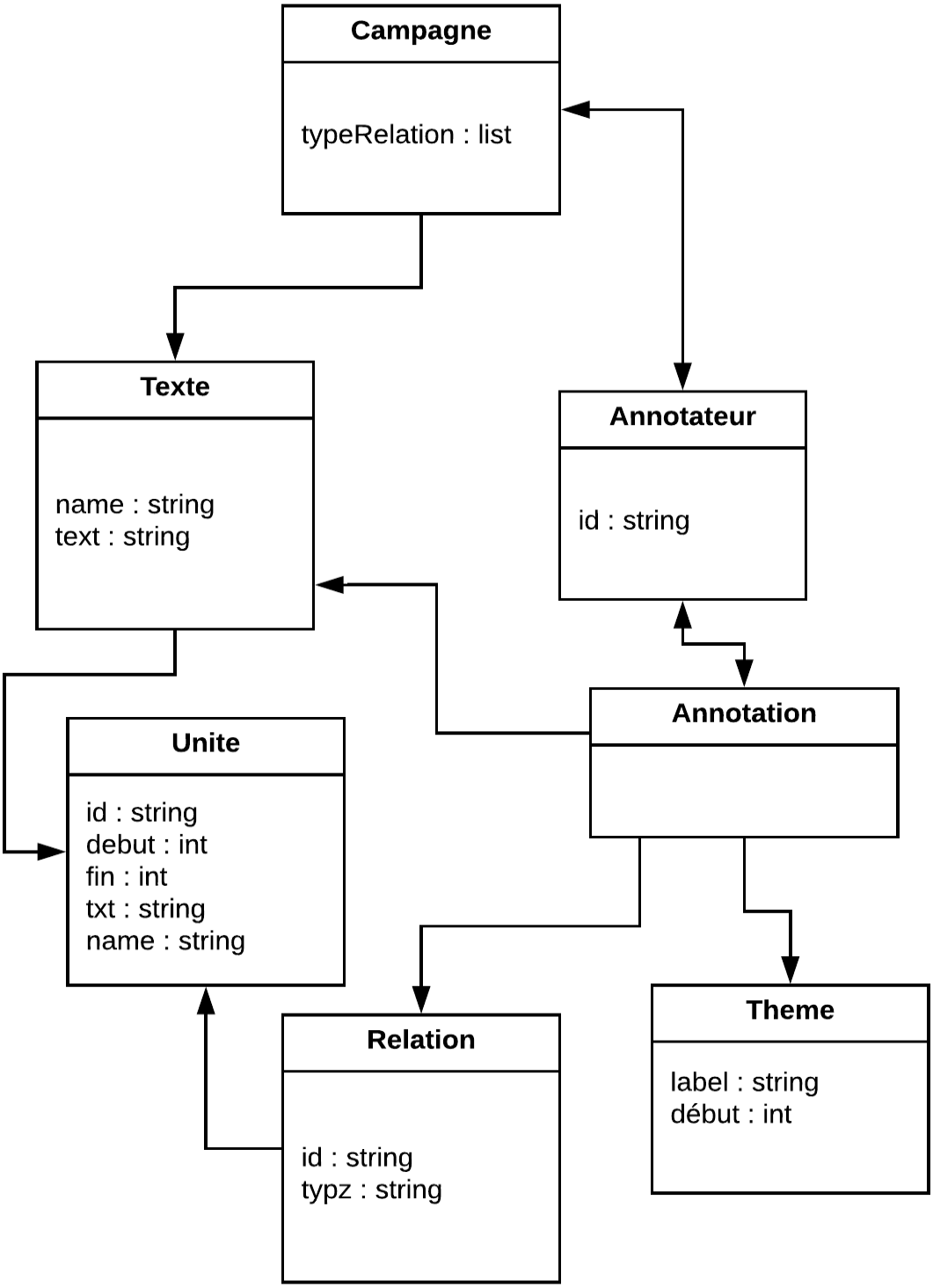
Figure 23 - dendrogramme Nord avec annotation d'expert

Annexe 5. CONSENSUS SUR CERTAINS TYPES DE RELATIONS : TABLEAU TEST DE WILCOXON-MANN-WHITNEY

Tableau 7 - Tests de Wilcoxon-Mann Whitney sur les consensus sur certains types de relations

	Bac_a_sable	Florence	Provocation	Nord	Concours	Sauveur
Florence	2.468554e-07	0.000000	0.000000	0.000000	0.000000	0.000000
Provocation	1.811769e-02	0.011939	0.000000	0.000000	0.000000	0.000000
Nord	1.141204e-02	0.022951	0.883486	0.000000	0.000000	0.000000
Concours	9.600735e-01	0.001389	0.336040	0.273110	0.000000	0.000000
Sauveur	1.717911e-04	0.328385	0.274216	0.379325	0.040797	0.000000
Volley	7.013044e-01	0.000036	0.099213	0.072461	0.714354	0.004136

Annexe 6. DIAGRAMME DES CLASSES



Annexe 7. LISTE DES LIVRABLES

Tous nos livrables sont disponibles sur notre Github à l'adresse suivante :

https://github.com/AlaixComet/AnnotationSlamPTut

Il contient notamment :

- *Campagne 2017 :* Les fichiers Glozz d'annotation pour la campagne précédente 2017
- *Campagne 2018* : Les fichiers Glozz d'annotation pour la campagne 2018
- *Arbres* : le dossier contenant les arbres générés pour chaque anontation
- *Glozz*-platform : Le logiciel Glozz et les fichiers nécessaires pour le paramétrer
- *Guide* : le guide d'annotation issu du projet précédent
- *Traitement données* : Le code python de traitement de données avec notamment :
  - *analyse.py :* nos scripts d'analyses
  - *parsing.*py : notre script de parcours des fichiers de campagne
  - *data.*py : notre structure de représentation des données
  - *randomizeList.*py : notre script de randomisation des listes d'annotation
  - *Analyse 2018.ipynb* : le Jupyter Notebook qui est l'équivalent de notre main
- *Données Annotation.xlsx :* le tableau des résultats des questionnaires de la campagne 2018
- *Installation notebook.pdf :* les instructions pour utiliser le notebook
- *Protocole d'utilisation de Glozz.pdf* : les instructions pour utiliser Glozz pour mener une campagne
- *Liste_Textes_Passation.csv* : l'ordre aléatoire que nous avons utilisé pour la campagne 2018