Projet tutoré

Trouble du langage et de la pensée : Campagne d'annotation

réalisé par :
Laurine Huber
Emilie Laurier

encadré par :
Maxime Amblard
Manuel Rebuschi

Master 1 Sciences de la cognition et applications - 2016/2017

Projet tutoré

Trouble du langage et de la pensée : Campagne d'annotation

étudiantes :
Laurine Huber
Emilie Laurier

tuteurs :
Maxime Amblard
Manuel Rebuschi

Master 1 Sciences de la cognition et applications - 2016/2017

Remerciements

Nous souhaitons dans un premier temps remercier Maxime Amblard et Manuel Rebuschi pour nous avoir suivies et guidées durant tout le projet. Ils nous ont permis d'avoir un travail organisé et nous ont appris les bonnes pratiques de la mise en place d'une telle campagne.

Merci à Valentin Steyer de nous avoir permis une meilleure compréhension de la plateforme Glozz. Ses réponses rapides à nos nombreux mails nous ont permis d'avancer rapidement sur toute la partie logicielle du projet.

Merci aussi aux divers intervenants des conférences auxquelles nous avons pu participer, celles-ci étaient très enrichissantes et nous ont permis de compléter notre compréhension de ce sujet vaste et complexe.

Nos derniers et plus précieux remerciements vont aux 46 annotateurs qui ont pris le temps de participer à cette campagne. Nous remercions tout d'abord nos proches, qui ont bien compris que leur aide était précieuse pour le bon déroulement de ce projet scientifique, mais aussi les personnes moins proches mais curieuses, qui se sont intéressées d'elles-même à notre travail.

Table des matières

1 Introduction du projet

1.1 Le contexte institutionnel

Dans le cadre de notre première année de Master en Sciences de la Cognition et Applications, nous avons réalisé un projet tutoré s'inscrivant dans l'équipe de recherche SLAM (Schizophrènes et Langage : Analyse et Modélisation) appuyée par la Maison des Sciences de l'Homme (MSH) de Lorraine et le Laboratoire Lorrain de Recherche en Informatiques et ses Applications (LORIA). Leur travail vise à étudier des conversations pathologiques, notamment celles impliquant des sujets atteints de schizophrénie. Leur approche est interdisciplinaire, elle mêle psychologie, linguistique informatique et philosophie. Une de leurs études vise à modéliser les entretiens entre des patients schizophrènes et des psychologues.

1.2 Le contexte scientifique

L'argument moteur de la démarche est basé sur un principe philosophique théorisé par les philosophes Quine et Davidson. Il s'agit du principe de charité [2], qui consiste a maximiser la vérité des croyances de l'autre en lui attribuant un maximum de rationalité. Dans le cadre de ce projet, nous partons c de la supposition que le contenu de pensée des schizophrènes est totalement conforme à notre logique. D'un point de vue sémantique le discours des schizophrènes est correct néanmoins il ne l'est pas d'un point de vue pragmatique.

Dans le but d'analyser ces troubles du langage et de la pensée, des corpus de conversations pathologiques ont dû être préalablement constitués à l'aide d'extraits de conversations obtenus lors d'entretiens entre psychologues et schizophrènes. Ils sont ensuite retranscrits en suivant une procédure de transcription [1]. Par exemple, la ponctuation est adaptée en fonction de la prosodie.

Une fois que le corpus est constitué, nous le regardons de différentes manières, avec des axes particuliers : un axe psychologique, un axe morpho-syntaxique et un axe sémantico-pragmatique. À travers ces axes, de multiples annotations sont possibles, mettant en avant différentes particularités du discours. Une annotation consiste à ajouter des informations à des données textuelles (ou autres), soit manuellement, soit par le biais d'un outil. Nous distinguons alors 3 types d'annotations :

- ne annotation de type psychologique permettant de mettre en avant des discontinuités apparaissant dans l'échange :

Ce type d'annotation est délicat car il requiert des annotateurs experts, capables d'identifier les ruptures dans le discours.

- Une annotation morpho-syntaxique sur l'ensemble du corpus :

Ce type d'annotation s'effectue par exemple à l'aide d'outils de tagging, en d'autres termes cela consiste à étiqueter un texte. Cette analyse avait été réalisée par d'anciens étudiants

1. Notation des unités phoniques du langage au m n de symboles et de signes graphiques conventionnels.

en projet tutoré, qui avaient montré qu'aucune observation n'était significative au niveau grammatical.

- Une annotation sémantico-pragmatique formelle, basée sur la S-DRT :

Ce type d'annotation est basé sur une théorie du discours, la DRT (Discourse Representation Theory), et plus précisément sur une extension de celle-ci, la S-DRT, un système de représentation formel qui permet de construire un arbre de relations discursives et conversationnelles. Ainsi la S-DRT permet de modéliser le discours. Cette théorie est née de deux grands courants de recherche, la sémantique dynamique et l'analyse du discours [1].

C'est sur ce dernier axe que notre projet tutoré se concentre, puisque nous tenterons à l'aide de la S-DRT, de repérer, les discontinuités dans le discours du schizophrène. Nous proposons de faire faire des annotations à des annotateurs non-experts, afin de réduire la subjectivité que pourrait engendrer l'annotation par des psychologues. En effet, les psychologues ont des informations que les annotateurs naïfs n'ont pas, qui peuvent influencer l'annotation et forcer l'apparition des résultats attendus.

Dans notre cas l'annotation comprend deux étapes. Elle consiste d'une part à lier des unités de textes par différents types de relations (l'aspect pragmatique). D'autre part elle consiste à repérer des thèmes, en d'autres termes les idées principales de la conversation (l'aspect sémantique).

1.3 L'objectif du projet

L'objectif de notre projet est de voir s'il [illegible] de repérer, à l'aide d'une méthode basée sur une annotation sémantico-pragmatique formelle, des discontinuités dans le discours du schizophrène et de supprimer la subjectivité des analyses d'annotateurs experts à l'aide d'un grande nombre d'annotateurs novices. Pour atteindre cet objectif nous avons dû mener à bien une campagne d'annotation en S-DRT. Une campagne d'annotation consiste à :

- Trouver les outils nécessaires pour la réaliser

Ne souhaitant pas axer l'ensemble de notre projet sur le développement d'un nouvel outil, nous avons choisi de mener la campagne avec la plate-forme Glozz, une plate-forme d'annotation développée par plusieurs équipes de recherche. Nous reviendrons dessus dans la section *Adaptation de Glozz*.

- Définir un guide d'annotation

Mettre en place un guide qui permette d'expliquer à l'annotateur le contexte du projet et la tâche qu'il doit accomplir sans pour autant influencer ses choix.

- Mener une pré-campagne

Tester les outils, le guide, et éventuellement récupérer les retours des annotateurs quant à la tâche demandée de façon à la modifier pour l'amélioration de la procédure.

- Recruter des annotateurs

Trouver un échantillon de personnes assez hétérogène pour mener une campagne dont les données extraites seront fiables : différentes catégories socio-professionnelles, âges, genre.

- Faire passer les annotations

Être disponible avant l'annotation et pendant la durée de celle-ci pour les manipulations techniques (changement de textes car le logiciel ne le permet pas automatiquement) et pour aider l'annotateur en cas de problème technique.

Après la réalisation de cette campagne, nous avons réalisé une analyse des données pour appréhender les annotations sous différents angles, permettant ou non, de mettre en avant des particularités au niveau du discours.

1.4 Le déroulement concret

Nous avons mis en place un calendrier (A.1) afin de nous organiser convenablement et de vérifier que les délais soient corrects ; autrement dit il fallait d'une part que la période de pré-campagne soit assez longue afin de pouvoir modifier le guide d'annotation selon les retours des pré-annotateurs et d'autre part une période assez longue pour le traitement des résultats. La campagne en elle-même est réalisée en seulement trois semaines. De ce fait nous avons planifié le programme suivant :

- Novembre à Janvier : lecture de la bibliographie
- Fevrier à Mars : pré-campagne (rédaction du guide, adaptation de Glozz)
- Avril : campagne d'annotations
- Mai : analyse des données et rédaction du rapport

En prime de ce programme, nous avons eu l'occasion de participer à la conférence "(In)cohérence du discours", durant laquelle des chercheurs ont présenté leurs avancées dans la modélisation du discours, dont ceux avec des locuteurs ayant une pathologie, telle que la schizophrénie. Nous avons également été conviées à une table ronde "Comprendre ou expliquer la folie ?", davantage axée sur les maladies mentales et notamment la schizophrénie, où M. Rebuschi y a présenté le projet SLAM.

Dans un premier temps nous présenterons notre travail relatif à la campagne d'annotation, la préparation et la réalisation de celle-ci. Dans un second temps nous aborderons l'analyse des données recueillies ainsi que les résultats obtenus. Pour finir, la dernière partie exposera les conclusions principales de notre étude.

2 Préparation de la campagne

2.1 Les éléments à préparer

Pour préparer la campagne nous nous sommes basées en grande partie sur le travail du projet tutoré de 2015. Néanmoins plusieurs points nous paraissaient primordiaux à améliorer pour simplifier l'annotation. Le premier concerne la rédaction du guide d'annotation, avec d'une part sa forme et d'autre part son contenu. Il y a deux ans, le guide de nos prédécesseurs faisait 8 pages et était peu synthétique. Par exemple, les relations rhétoriques s'étendaient sur 3 pages en étant listées les unes après les autres. Nous voulions pour réaliser notre campagne un guide simplifié, ergonomique et esthétique visuellement. C'est-à-dire compact, adapté pour n'importe quel annotateur et lisible. Ce point nous paraissait primordial pour faire passer l'annotation pour quelque chose de moins formel que cela en a l'air, et ainsi obtenir le plus grand nombre d'annotations possibles. Dès lors, nous avons élaboré plusieurs versions du guide avant d'arriver à un résultat satisfaisant, nous détaillerons ces versions par la suite.

Concernant le contenu du guide, et plus précisément les relations rhétoriques, il fallait savoir que l'ancien projet tutoré comptait quinze relations. Parmi ces relations il y avait des ambiguïtés (descriptions équivoques, difficiles à appréhender) que nous avons voulu réduire au maximum, toujours dans le but de simplifier la tâche et de ne pas décourager l'annotateur. Nous développerons ultérieurement une partie sur ces relations.

Enfin le dernier point sur lequel nous avons travaillé pour simplifier l'annotation concerne la segmentation préalable des textes en unités sur le logiciel Glozz. En effet nos prédécesseurs ont souligné que l'annotation dans sa globalité prenait du temps et pouvait être laborieuse pour les annotateurs. Même si un léger biais est intégré en agissant de la sorte, cela permettait d'enlever une tâche complexe à l'annotateur, et ainsi de réduire la durée de l'annotation et de toucher plus de volontaires. Nous reviendrons sur ce point dans la partie consacrée à l'adaptation du logiciel.

2.2 Rédaction du guide

2.2.1 La forme du guide

Nous avons élaboré 3 versions du guide avant d'arriver à un résultat satisfaisant. Entre chaque version du guide un rendez-vous de discussion était organisé avec nos tuteurs pour réfléchir ensemble aux améliorations possibles et aux points potentiellement problématiques. La première version du guide comprenait :

- un guide formel avec l'introduction du projet suivie de la mission à effectuer
- l'explication de la prise en main de Glozz
- un tableau expliquant les différentes relations rhétoriques.
- une fiche de synthèse résumant les différentes relations rhétoriques afin que l'annotateur l'ait sous les yeux, sans avoir à retourner dans les explications du guide.

La diversité des représentations des relations rhétoriques (le tableau du guide et la fiche de synthèse) pouvait rendre confus l'annotateur.

Cela nous a amenées à faire une seconde version du guide, où nous avons retiré le tableau des relations rhétoriques, pour lister une série d'exemples en plus de ceux présents sur la fiche de synthèse, afin de d'aider l'annotateur à comprendre les notions. Toutefois ce n'était plus la diversité des représentations qui posait problème, mais celle des exemples, qui amenait l'annotateur à jongler entre le guide formel et la fiche de synthèse.

Dès lors nous avons abouti à une troisième et dernière version du guide, en incorporant directement la fiche de synthèse dans le guide. La version définitive du guide d'annotation se présente sous forme de livret composé de quatre pages.

- la page de couverture comprend l'introduction du projet ainsi que la mission à effectuer
- la double page contient la fiche de synthèse
- le dos du livret contient le tableau explicatif de la prise en main de Glozz

Cette version est devenue définitive suite à la pré-campagne, après validation de nos tuteurs.

A la suite de ces discussions nous avons donc revu la forme du guide et également son contenu, comme par exemple le nombre de relations rhétoriques.

2.2.2 Le contenu du guide

Le groupe du projet tutoré de 2015 avait 15 relations rhétoriques dans son guide :

Narration	Elaboration	Question
Réponse	Elaboration : Explication	Question : Conduite
Réponse phatique	Elaboration : Prescription	Méta-question
Suite d'élaboration	Evaluation	Demande d'élaboration
Conduite	Phatique	Contre-élaboration

TABLE 1 – Tableau des anciennes relations rhétoriques

Nous avons jugé que certaines des relations rhétoriques risquaient de rendre confus l'annotateur "novice" à cause d'une finesse analytique difficile à appréhender. De ce fait nous avons décidé de restreindre la diversité des relations utilisables, afin de rendre l'annotation plus efficace. Nous avons par exemple fusionné *Elaboration* et *Élaboration : Explication* en une même relation : *Elaboration descriptive*. Nous avons supprimé les relations *Suite d'élaboration*. *Réponse phatique, Demande d'élaboration*, et *Question conduite* car leur définition était similaire à des relations déjà présentes. Ainsi, nous comptons au final dans notre guide seulement 10 relations nous permettant tout de même de bien annoter le discou

Narration	Elaboration Descriptive	Elaboration prescritpive	/
Elaboration évaluative	Contre-élaboration	Conduite	/
Phatique	Question	Réponse	Méta-question

TABLE 2 – Tableau de nos relations rhétoriques

Nous avons regroupé ces relations dans le guide en les associant à une couleur. En violet nous avons mis Narration, Elaboration descriptive, Elaboration prescriptive et Élaboration évaluative ; car l'idée générale est l'élaboration. En rouge Conduite et Phatique, car ces deux relations se rapportent plus à la conversation qu'au contenu de celle-ci. En bleu Contre-élaboration. En vert, Réponse, Question, et Méta-question. Nous retrouvons ces mêmes relations et couleurs associées dans Glozz (voir Figure 1). Pour plus d'explications, nous avons mis en annexe la définition de chacune de ces relations ainsi que des exemples les illustrant.

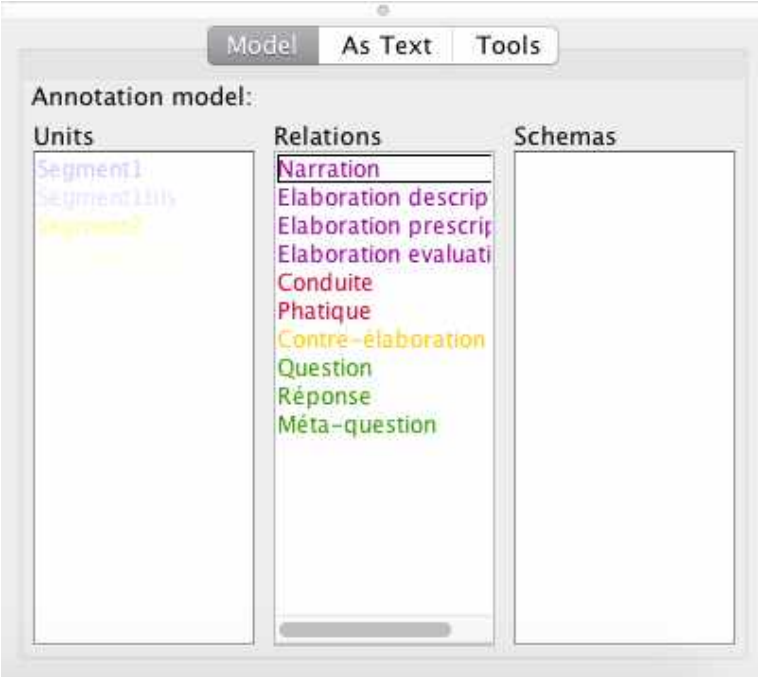


FIGURE 1 – Relations dans Glozz

2.3 Le choix des textes

Nos encadrants nous ont choisi 3 textes ( ) à analyser. Nous avons deux extraits de dialogues entre un psychologue et un schizophrène : "Les deux Florence" (contenant 14 unités) et "Pro-vocation" (23 unités) qui contiennent tous deux des ruptures discursives selon les annotations d'experts. Nous avons ajouté un troisième texte avec l'accord de nos tuteurs, "Nord1" (13 unités) comportant également une rupture discursive, plus subtile car le patient en question est suivi et traité. Enfin suite à la pré-campagne un quatrième texte s'est ajouté : le texte "Bac-à-sable" (8 unités) que nous avons rédigé nous-mêmes. Il permet d'accompagner les annotateurs sur la prise en main de Glozz et la compréhension de la tâche avant de commencer l'annotation des textes plus complexes.

2.4 Adaptation de Glozz

L'adaptation de Glozz se déroule en trois étapes. Il faut dans un premier temps réfléchir à la forme que l'on veut donner à l'outil , puis préparer l'outil, pour ensuite arranger et intégrer les données qui seront annotées.

En réfléchissant aux attentes que pourraient avoir les annotateurs quant au déroulement de la mission, plusieurs choix se sont présentés à nous et nous avons donc fait plusieurs modifications par rapport à l'ancien projet tutoré. Pour plus de clarté pour l'annotateur, nous faisons correspondre les relations du guide avec les relations de Glozz par un système de couleurs. Cela permet d'enlever une subtilité à la partie technique de l'annotateur car il trouve rapidement la relation dans Glozz après l'avoir repérée dans le guide.

Sur la suggestion de notre encadrant Maxime Amblard, nous pré-découpons préalablement les phrases en unités pour deux raisons : la première est le souhait d'alléger la tâche de l'annotateur, lui permettant ainsi de se concentrer pleinement sur la mise en place des relations et des thèmes. Aussi, cela nous permet d'avoir une base similaire pour toutes les annotations, nous permettant donc plus de facilité lors du traitement des données, et notamment de la comparaison des arbres. Nous en parlerons plus en détails dans la section analyse de données.

En plus du découpage, nous ajoutons des couleurs aux unités. Le bleu et le jaune sont utilisés pour distinguer les interlocuteurs, et une nuance de celles-ci est utilisée pour distinguer les différentes unités.

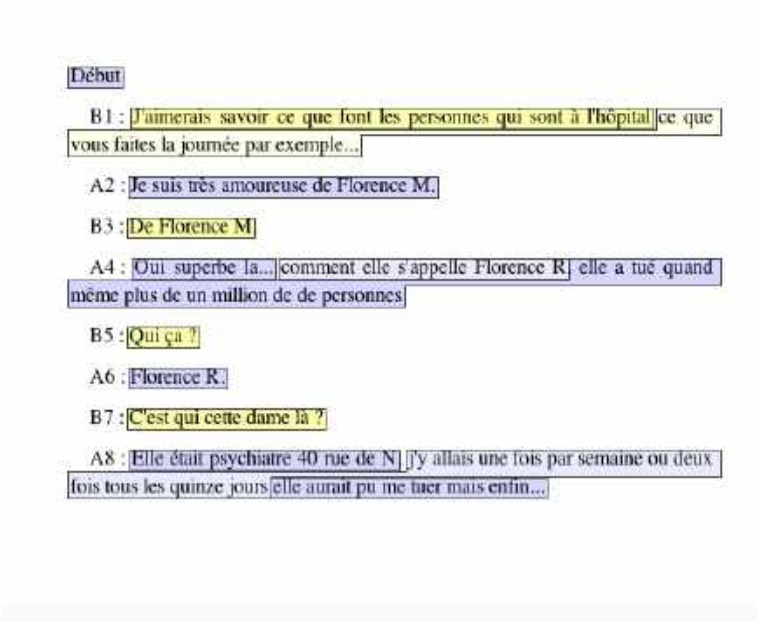


FIGURE 2 – Illustration du découpage des unités

Nous modifions aussi le système de mise en place de thèmes. L'ancien groupe obligeait l'annotateur à donner un thème pour chaque unité. S'ils ne le faisaient pas, par paresse ou par manque de temps, cela pouvait avoir une influence négative sur l'analyse des données. Pour se dédouaner de cela, nous n'obligeons pas l'annotateur à définir le thème de chaque

unité, mais nous lui demandons de mettre un drapeau définissant le thème, à chaque fois qu'il estime apercevoir un changement pertinent dans la discussion. Cela nous permet aussi d'éviter la redondance : si 5 unités consécutives parlent de nourriture, le thème "nourriture" apparaîtra une seule fois, au début de la première unité, et le changement de thème sera signalé par un nouveau thème après la dernière unité.

La deuxième démarche consiste en l'adaptation de Glozz aux besoins préalablement établis ci-dessus. La mise en place interne de Glozz et notamment la définition des relations et des unités se fait grâce à des fichiers XML spécifiques au logiciel. Ce fichier permet de définir le nom, la couleur d'arrière plan et la couleur de texte d'une relation ou d'une unité. Les balises "unit-style" et "relation-style" permettent de définir s'il s'agit d'une unité ou d'une relation. La balise "background-color" définit la couleur d'arrière plan, et "line-color" la couleur des flèches et du texte pour les relations. "invisibility value" est tout le temps à false pour qu'aucune unité ou relation ne soit cachée. Pour les unités, nous avons donc 4 types différents, une paire bleu foncé / bleu clair, pour l'un des locuteurs, une paire jaune foncé / jaune clair pour le deuxième locuteur. Pour les unités, le nom n'est pas important puisque les annotateurs n'ont en aucun cas besoin de s'en servir, le découpage en unités étant fait par nous même. Voici un aperçu du fichier concernant la partie des unités.

```
<unit-style>
<type name="Segment1"/>
<background-color b="255" g="153" r="153"/>
<invisibility value="false"/>
</unit-style>
<unit-style>
<type name="Segment1bis"/>
<background-color b="255" g="204" r="204"/>
<invisibility value="false"/>
</unit-style>
<unit-style>
<type name="Segment2"/>
<background-color b="51" g="255" r="255"/>
<invisibility value="false"/>
</unit-style>
<unit-style>
<type name="Segment2bis"/>
<background-color b="204" g="255" r="255"/>
<invisibility value="false"/>
</unit-style>
```

FIGURE 3 – code pour les unités

Pour les relations, le principe est le même mais ici le nom est important puisqu'il définit le type de relation que l'annotateur choisit d'utiliser. Ainsi dans ce fichier, nous définissons toutes les relations du guide et y associons la couleur utilisée dans le guide, comme dit précédemment. Voici un exemple :

```
<relation-style>
<type name="Conduite"/>
<line-color b="51" g="0" r="204"/>
<background-color b="0" g="0" r="255"/>
<invisibility value="false"/>
</relation-style>
<relation-style>
<type name="Contre-élaboration"/>
<line-color b="0" g="204" r="255"/>
<background-color b="0" g="0" r="255"/>
<invisibility value="false"/>
</relation-style>
<relation-style>
<type name="Elaboration descriptive"/>
<line-color b="153" g="0" r="153"/>
<background-color b="0" g="0" r="255"/>
<invisibility value="false"/>
</relation-style>
```

FIGURE 4 – cod pour les relations

Une fois ce fichier XML définit et chargé dans Glozz, nos relations et nos différents types d'unités apparaissent dans le logiciel.

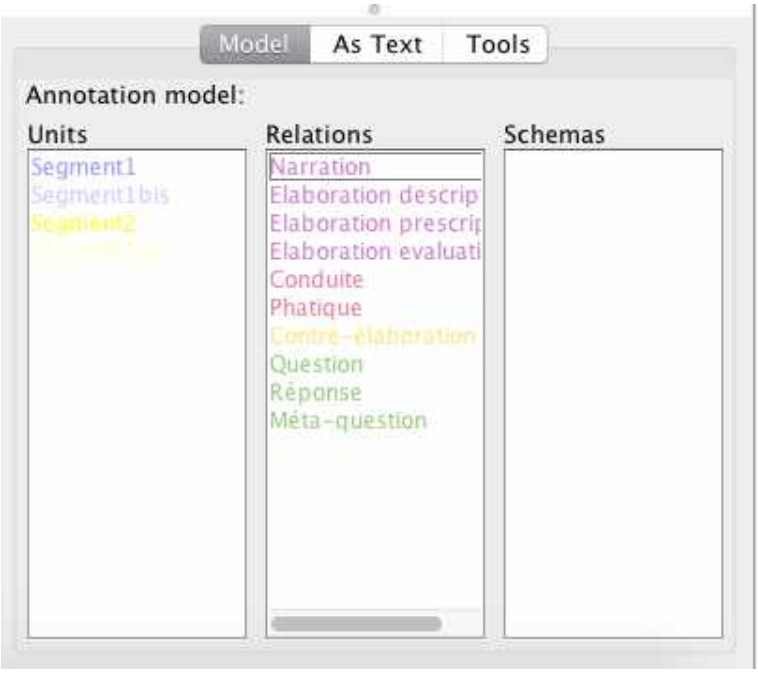


FIGURE 5 – Unités et Relations dans Glozz

Une fois l'outil prêt, il a fallu préparer et intégrer les textes choisis dans Glozz, et les découper en unités comme énoncé précédemment. Dans un premier temps le texte est découpé en unités à la main, de façon à vérifier la cohérence du découpage avec nos tuteurs. Une fois cela revu, corrigé, et validé, nous le copions dans un fichier .txt en respectant le format énoncé ci-dessus. Ainsi, une ligne correspond à une unité. Nous ajoutons une unité appelée "Déb en première ligne de texte, pour forcer la racine de l'arbre [2]. Une fois les textes prêts, nous les importons dans Glozz et les découpons en unités à la main. Glozz nous permet d'enregistrer puis d'importer un découpage en unités sous forme de fichier .csv, ce que nous faisons pour réutiliser ces découpages pour chaque annotateur.

Enfin, afin que notre adaptation de Glozz puisse être réutilisée pour d'autres annotations nous avons rédigé un tutoriel [3], avec des captures d'écrans, qui reprend pas à pas les procédures nécessaires pour créer et ajouter les fichiers essentiels, charger un texte, l'annoter et enregistrer le tout ; afin que tout ceci puisse être utilisé à la suite du projet.

2. Chaque annotateur est incité à relier la première unité à début en respectant une convention. Cela est nécessaire pour avoir une racine similaire pour chaque arbre.

3. Le tutoriel n'est pas annexe mais en livrable compte tenu de son nombre conséquent de pages

2.5 Pré-campagne

Une fois le guide rédigé et Glozz adapté pour nos annotations nous avons pu mettre en place une pré-campagne. Celle-ci nous a permis de tester notre guide, ainsi que les textes découpés à annoter. Nous avions choisi de faire annoter les textes dans l'ordre suivant "Provocation", "Les deux Florence", "Nord1" à quatre pré-annotateurs. Nous avons choisi l'ordre suivant pour une question de longueur de textes, car l'extrait "Pro-vocation" est le plus long, nous l'avons mis en premier pour ne pas décourager les annotateurs, afin qu'ils puissent accomplir l'annotation entièrement. Quant au nombre de pré-annotateurs nous en avons choisi 4 de notre filière par facilité de contact et pour solliciter le reste de notre promo pour la campagne. Leur mission était la suivante : il fallait dans un premier temps relier les unités à l'aide des relations rhétoriques puis il fallait donner un ou des thèmes à la conversation.

A l'issue de ces tests nous avons constaté que les annotations prenaient 30 à 45 minutes par annotateur. La tâche nécessitait du temps, toutefois aucun des pré-annotateurs n'a trouvé la tâche rebutante. Lors de cette pré-campagne nous avons tenu compte des remarques des pré-annotateurs afin d'effectuer les améliorations nécessaires pour la campagne. Plusieurs points ont été soulevés :

- Nous faisions commencer l'annotation par le texte le plus long, car nous pensions que ce serait moins décourageant pour les pré-annotateurs de finir avec des textes beaucoup plus courts. Néanmoins le fait de débuter avec un texte long, ne facilitait pas la prise en main du logiciel et du guide. La difficulté de compréhension du texte apparaissait comme une contrainte en plus, qui rendait la tâche plus compliquée que prévu. Suite à cela, nous avons alors décidé non seulement de commencer la campagne avec un texte plus court, mais aussi de créer un texte " bac-à-sable " qui nous permette de conseiller et guider l'annotateur dans la première démarche de prise en main du logiciel, et de le laisser ensuite seul pour les 3 autres textes qui constituent nos réelles données.

- Il est apparu que souvent les pré-annotateurs ne comprenaient pas bien comment annoter les thèmes puisqu'ils avaient tendance à mettre des annotations de thèmes partout. Il a alors fallu que nous ajoutions un point dans le guide pour donner un nombre maximum de thèmes. Nous avons fixé 10 thèmes par texte pour plus de simplicité et de compréhension. Nous avons choisi ce maximum car le texte le plus court contient 8 unités et le plus long en contient 23, 10 thèmes laissant à l'annotateur un choix assez large pour le texte le plus long sans l'influencer ou le priver.

- Enfin dernière remarque et pas des moindres, initialement nous n'avions pas touché à la structure du dialogue des fichiers txt que nous utilisions sous Glozz, lorsque nous différencions les unités au sein d'une même intervention, nous alternions avec deux nuances d'une même couleur sans retour à ligne. Ceci a gêné un des pré-annotateurs qui ne discernait pas bien les différentes unités pour un même locuteur, car dans le guide nous avions mis chaque unité à la ligne pour chaque locuteur. Ainsi le pré-annotateur oubliait de relier des unités pensant qu'il y avait qu'une unité alors qu'en réalité il y en avait par exemple quatre. Suite à cette remarque, nous avons décidé de reprendre chaque fichier txt, et de mettre chaque unité à la ligne afin que les annotateurs aient la même chose sur la plateforme de Glozz ainsi que sur la guide, en vue d'enlever toute ambiguïtés.

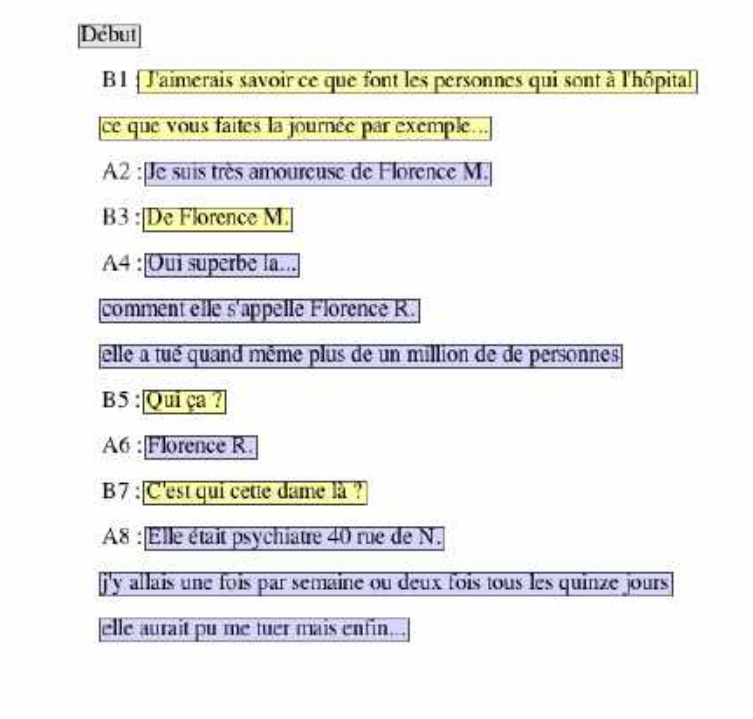


FIGURE 6 – Découpage des unités suite à la pré-campagne

Cette pré-campagne a été constructive, elle nous a permis d'apporter les ajustements nécessaires, de plus, les données recueillies lors de celle-ci nous ont servi pour commencer les procédures d'analyses de données, en d'autres mots de tester les scripts de visualisation d'arbre ou d'extraction de données.

3 Réalisation de la campagne

3.1 Recrutement

Pour des questions d'organisation, nous avons décidé de réaliser les annotations sur plusieurs sites distincts : Nancy, Bourgogne, Alsace. Nous avons aussi décidé de concentrer les annotations sur 3 semaines, de façon à avoir davantage de temps avant de réaliser la campagne, de la préparer et après celle-ci d'analyser les données recueillies.

Nous avons dans un premier temps cherché dans nos relations quelles personnes pourraient être disponibles durant nos sessions pour réaliser l'annotation. Il a aussi fallu que nous réfléchissions aux âges et catégories socio-professionnelles de nos proches de façon à sélectionner des échantillons les plus hétérogènes possibles de façon à ne pas fausser l'analyse finale. Une autre contrainte était de ne choisir que des personnes majeures.

Dans un second temps, nous avons créé un formulaire d'inscription (un Go  Form) pour les personnes extérieures à notre entourage, de façon à récupérer un maximum d'intéressés. Nous avons envoyé ce formulaire d'inscription sur les réseaux sociaux et avons eu seulement 4 réponses. Le recrutement sur notre entourage a été plus efficace que celui via le formulaire, nous avions anticipé cela en nous imposant un objectif de 20 annotations pour chacune de nous dans nos régions respectives. Toutefois nous ne nous sommes pas contentées des réponses du formulaire, nous avons recruté de nous mêmes d'autres personnes sur Nancy.

3.2 Déroulement

Pour chaque annotateur, nous leur donnions le guide explicatif à lire avec soin, puis après la lecture nous l'aidions à prendre en main Glozz avec le texte "bac à sable". Une fois que l'annotateur était prêt et que tout était clair pour lui, nous le laissions en autonomie sur 3 extraits à annoter, dans l'ordre suivant "Les deux Florence", "Pro-vocation", "Nord1". Nous intervenions seulement pour des problèmes techniques avec Glozz, ainsi que pour enregistrer leurs annotations et leur présenter le texte suivant.

A la fin des annotations, nous faisions passer à l'annotateur un questionnaire, nous permettant de recueillir des informations concernant leur âge et leur situation socioprofessionnelle ainsi que leur ressenti vis-à-vis de l'annotation (compréhension, durée, et prise en main de Glozz). Ces données nous permettent de comparer leur ressenti avec ce qu'ils ont réellement effectué. Les annotations duraient en moyenne 30 à 45 minutes par annotateur, et nous n'avons eu aucun abandon, ni d'annotation partielle.

Au final nous avons eu 18 annotations en Alsace, 17 annotations à Auxerre, 11 annotations sur Nancy soit un total de 46 annotations. Nous détaillerons le profil des annotateurs dans la partie Analyse des données.

3.3 Amélioration possibles

A la fin de notre campagne nous avons réfléchi aux améliorations possibles. D'une part notre questionnaire suite aux annotations, nécessitait d'être approfondi, nous le trouvions au final un peu léger. D'autre part, la diffusion du Google Form d'inscription est à revoir, étant donné que nous avons eu peu de retours, il aurait fallu le diffuser probablement plus tôt et le relancer hebdomadairement pour obtenir plus d'annotateurs extérieurs.

4 Analyse des données

4.1 Hypothèses

Plusieurs informations sont pertinentes à analyser du point de vue de ce projet.

Dans un premier temps, s'agissant d'une campagne faisant intervenir des annotateurs externes, il fallait vérifier l'hétérogénéité des profils des annotateurs. Nous avions essayé de choisir des candidats de façon réfléchie pour influencer ce point, mais il est aussi nécessaire de le vérifier après recueil des données, pour éviter tout biais au niveau de la compréhension de la mission. Par exemple, des annotateurs provenant de la même filière vont possiblement être amenés à avoir une certaine compréhension bien particulière de la mission, et ainsi influencer l'analyse en ayant tous la même réflexion. Il était donc nécessaire, pour une analyse des données pertinente, de vérifier que notre échantillon soit suffisamment hétérogène pour éviter ce genre de biais au maximum.

Ensuite, dans le but de vérifier la consistance des données et surtout d'avoir un premier aperçu de nos annotations, il fallait réaliser des analyses simples comme la distribution et la fréquence d'apparition des relations ou des thèmes. Pour cela il a fallu visualiser les données sous différents points de vue de façon à essayer de montrer, ou non, un début d'analyse. Par exemple, en regardant les fréquences d'apparition des différentes relations à travers les textes, nous pouvons montrer, ou non, si une certaine relation semble être plus utilisée qu'une autre.

La dernière analyse consiste en l'observation des arbres d'annotations, et est à réaliser si les données précédentes sont pertinentes. En effet, si l'on arrive à montrer avec les deux premières analyses, que l'échantillon ou bien les données d'annotations sont biaisées, la dernière étude ne sera pas autant significative que si l'on arrive à montrer les deux premiers points. Cependant, nous avons tout de même décidé d'y réfléchir et d'implémenter une version d'un algorithme permettant de trouver, si oui ou non, une rupture conversationnelle existe dans un texte.

Dans le but d'obtenir des informations sur l'annotation à proprement parler ainsi que sur son interprétation, 2 points étaient importants :

- Savoir si plusieurs arbres distincts étaient produits par l'annotateur sur un seul et même texte :

Dans ce cas, 2 ou plusieurs sous-groupes de conversations apparaissent laissant suggérer une rupture discursive.

- Vérifier l'ordre chronologique d'apparitions des unités lors d'un parcours préfixe de l'arbre issu de l'annotation :

En effet, en réalisant un arbre correspondant à une conversation bien construite, on remarque que son parcours préfixe correspond à l'ordre d'apparition des segments/unités de phrase dans le texte. Ainsi, si l'on arrive à trouver une cassure dans cet ordre, celle-ci sera représentative de l'endroit d'une rupture dans le texte. En d'autres termes, l'interlocuteur intervenant

à ce moment précis du discours, fait référence à un élément différent de celui auquel un interlocuteur typique référerait.

4.2 Illustration des ruptures possibles

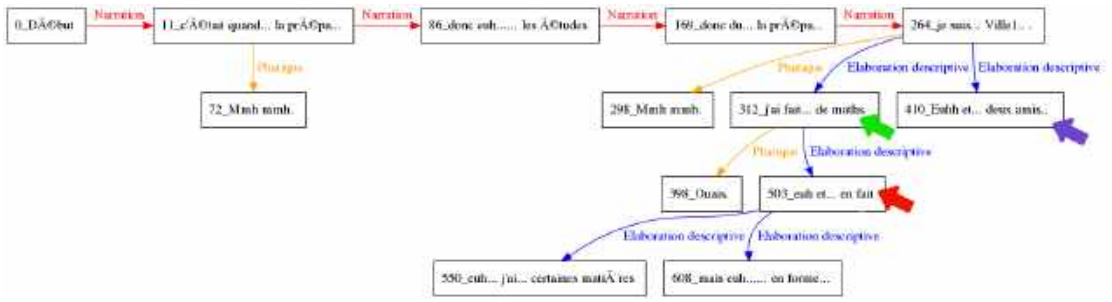


FIGURE 7 – Arbre du texte "Nord1" - annotateur expert

Ici, à cause du noeud 410 (flèche bleue), l'unité 503 (flèche rouge) ne peut pas se rattacher à l'unité 412 (flèche verte). Ainsi, le locuteur réfère à un élément de la conversation auquel il n'était pas censé référer s'il respectait une bonne structure conversationnelle. Ce type de rupture est appelé "rupture de la frontière droite", faisant référence à l'accessibilité des unités de conversations, et ce type de comportement conversationnel correspond donc à une rupture.

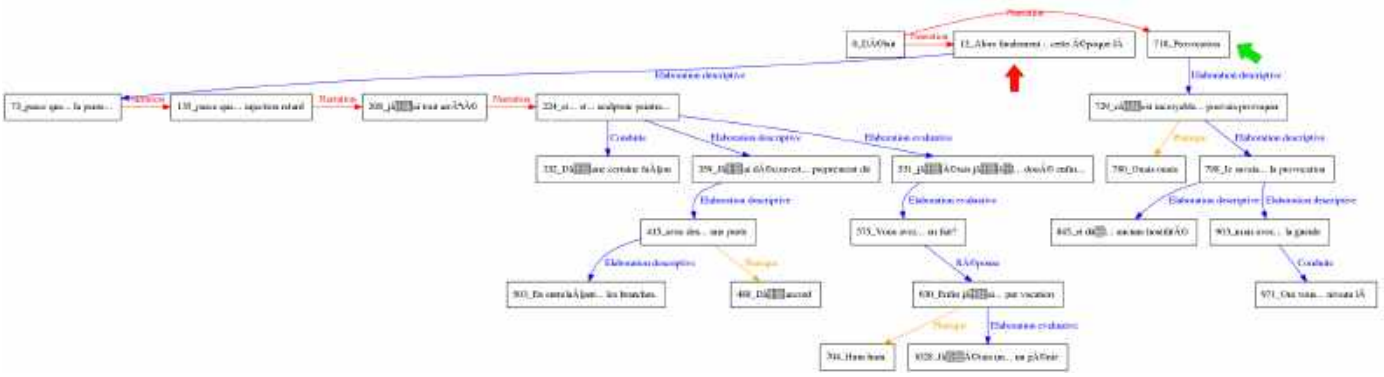


FIGURE 8 – Arbre du texte "Provocation" - annotateur expert

Ici, les flèches verte et rouge indiquent la racine de 2 sous-arbres issus d'une même annotation. Cela est aussi démonstrateur d'une rupture puisque cela signifie que l'annotateur, ne sachant pas ou relier une unité, l'a reliée au début. Cela signifie donc qu'une unité de texte ne référait en rien au reste de la conversation, suggérant alors une mauvaise interprétation de la part du locuteur.

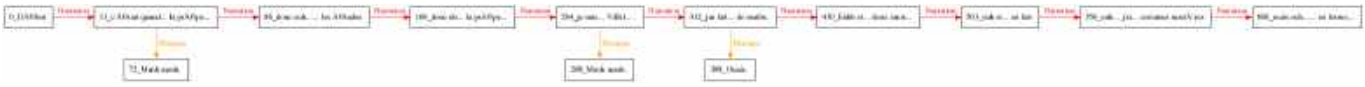


FIGURE 9 – Arbre du texte "Nord1" - annotateur novice

Dans cette annotation, le parcours des unités reliées suit bien l'ordre d'apparition des unités dans le texte, et aucun sous arbre ne découle de la racine, cette interprétation laisse donc suggérer qu'il n'y a pas de rupture dans cette conversation.

4.3 Description des données recueillies

Dans le but de vérifier la première condition, qui est l'hétérogénéité de notre échantillon d'annotateurs, nous avons décidé de faire passer à chacun un questionnaire très rapide comme introduit précédemment. Nous y avons recueilli leur âge, leur catégorie socioprofessionnelle, et leurs retours sur 3 points de la campagne : la compréhension de la campagne, la compréhension du logiciel, et la durée de l'annotation. Cela nous permet par ailleurs de mettre potentiellement en parallèle de "mauvaises [4]" annotations avec des profils de personnes ayant mal compris la mission, le logiciel, ou alors des personnes ayant trouvé l'annotation trop longue qui se seraient de ce fait dissipées.

Ces données sont recueillies grâce à un Google Form que nous faisions passer en fin de campagne à nos annotateurs. Un identifiant est attribué à l'annotateur au début de l'annotation, et est demandé lors du questionnaire. Cela nous permet de garder l'anonymat des annotateurs mais aussi de pouvoir mettre en parallèle les données issues du formulaire avec les données issues de l'annotation. De plus, les annotateurs issus de la campagne d'Emilie sont représentés sous la forme AXX et celles de Laurine sous la forme BXX. Cela nous permet de vérifier, si besoin, si la personne chargée de faire passer l'annotation ne produit pas une influence sur l'annotation.

Lors de la procédure d'analyse, le fichier csv issu du Google Form est analysé de façon à extraire les données correspondantes à l'annotateur traité, et à les ajouter au modèle contenant les données issues de l'annotation de celui-ci. Nous en parlerons plus en détails dans la section suivante.

Les données issue de l'annotation du texte sont quant à elles enregistrées sur nos ordinateurs respectifs grâce au logiciel Glozz introduit précédemment. Tout comme les données de configuration de Glozz, les données produites par l'annotation sont exportées au format XML. Nous nous partageons les fichiers au fur et à mesure des annotations via un dossier Google Drive.

4. Par mauvaise, nous entendons faussée. Une telle annotation se manifeste si elle sort du lot, c'est-à-dire, si la personne a employé significativement plus souvent un certain type de relation, si les thèmes qu'elle a définis ne sont en accord avec aucun des autres thèmes etc.

Comme évoqué précédemment, ce qui nous intéresse dans l'analyse, sont les textes, le découpage en unités, les relations, et les thèmes. Les textes sont définis dans des fichiers .txt avant même l'annotation. Le reste est obtenu après l'export CSV du fichier XML généré par Glozz. Les relations sont encadrés de balises <relation>, les unités par des balises <comment> [5], et les unités par des balises <unit>. Ci-dessous, un extrait de fichier XML, illustrant le format des balises citées :

```
<relation id="lhuber_1492009518309">
<metadata>
<author>lhuber</author>
<creation-date>1492009518309</creation-date>
<lastModifier>n/a</lastModifier>
<lastModificationDate>0</lastModificationDate>
</metadata>
<characterisation>
<type>Elaboration evaluative</type>
<featureSet/>
</characterisation>
<positioning>
<term id="vsteyer_1492008998282"/>
<term id="vsteyer_1492008998266"/>
</positioning>
</relation>
```

```
<unit id="TXT_IMPORTER_1492008986324">
<metadata>
<author>TXT_IMPORTER</author>
<creation-date>1492008986324</creation-date>
<lastModifier>n/a</lastModifier>
<lastModificationDate>0</lastModificationDate>
</metadata>
<characterisation>
<type>paragraph</type>
<featureSet/>
</characterisation>
<positioning>
<start>
<singlePosition index="0"/>
</start>
<end>
<singlePosition index="5"/>
</end>
</positioning>
</unit>
```

```
<comment>Psychiatrie</comment>
```

FIGURE 10 – Extraits de fichier XML illustrant les balises

Concernant l'analyse des arbres, c'est à partir d'un fichier .dot [6] que nous pouvons faire nos traitements et appliquer nos algorithmes de détection de rupture. En effet, le groupe précédent avait développé un outil en java permettant, à partir d'un fichier XML issu de Glozz, de créer l'arbre dans un fichier .dot, pouvant quant à lui être convertit en .png [7] pour la visualisation. Nous nous sommes servi de cet outil pour créer tous les .dot correspondant aux diverses annotations. Dans un premier temps, nous avons développé l'outil permettant de détecter si des ruptures sont présentes dans le texte. Pour cela, nous parcourons toutes les relations du .dot : si lors du parcours, un des noeuds père n'est pas déjà présent dans la liste des noeuds déjà parcourus, cela veut dire que ce noeud va devenir la racine d'un nouvel arbre.

5. Dans le logiciel Glozz, nous avons choisi de faire définir les thèmes sous forme de commentaire. Ainsi l'annotateur plaçait le commentaire où il le souhaitait, et donnait comme texte des mots clés définissant le thème abordé à ce moment du texte.

6. Un fichier dot est un fichier décrivant un graphe dans le language DOT, language qui décrit des graphes orientés ou non orientés à l'aide de noeuds et de flèches définissant les arcs.

7. Le png est un format d'image. divers outils en ligne de commande permettent à partir de la description .dot d'un fichier, d'obtenir sa représentation graphique sous forme d'image.

Ainsi, nous pouvons extraire les annotations qui génèrent des représentations mal formées, et faire une étude sur la fréquence d'apparition de celles-ci. Les résultats se trouvent dans la partie analyse des arbres. Aussi, à partir de ce .dot, nous pouvons faire abstraction du texte à proprement parler et traiter nos données comme de simples éléments identifiés par un unique. Pour détecter les ruptures à l'intérieur des arbres, nous avons alors développé en python un outil permettant d'analyser ce fichier .dot pour transformer sa structure en une structure exploitable par python : une simple liste. Un exemple d'arbre et sa représentation sous forme de liste python :

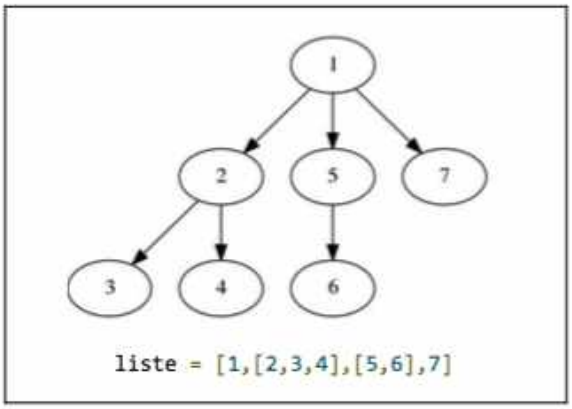


FIGURE 11 – Arbre et représentation sous forme de liste python

Une fois un .dot convertit en cette structure, il est possible de développer l'algorithme permettant de trouver les ruptures. Dans notre cas un nœud est défini par l'identifiant [8] issu du .dot qui correspond à une unité du texte. Il est associé à une chaîne de caractère défini par X label, "X" étant le chiffre correspondant à l'index du début de l'unité dans le texte d'origine, et "label" correspondant au texte de l'unité. Nous souhaitons donc à l'aide de notre algorithme, avoir un aperçu de la distribution de l'apparition de ces ruptures dans l'ensemble de notre échantillon d'annotations.

4.4 Technologies et procédures d'analyses

Pour ce qui est de l'analyse des résultats à proprement parler, c'est à dire des arbres issus de l'annotation, ainsi que des relations et des thèmes définis, nous avons besoin d'analyser les fichiers XML générés par le logiciel Glozz. Pour mettre en oeuvre nos notions de programmation acquises durant notre formation, nous avons choisi de traiter ces données à l'aide de python.

Plusieurs outils complémentaires à python nous permettent de travailler rapidement. Dans un premier temps, c'est la librairie "lxml" qui nous permet de d'analyser rapidement les fichiers générés par Glozz, et de les intégrer à des structures de données exploitables facilement.

8. Cet identifiant est donné aléatoirement lors de la création du .dot, donc dans le code java

Une fois ces données extraites, nous avons choisi de les analyser à l'aide de la librairie "pandas". Cette librairie complémentaire à python permet de structurer les données sous forme de tableaux indexés, et ainsi de pouvoir les analyser rapidement, à l'aide de filtres permettant une sélection restreinte. En effet, tout comme les requêtes en base de données, pandas permet de sélectionner des éléments en fonction d'une condition. Cela nous permet différentes approches puisqu'il devient alors facile de choisir sur quel texte nous voulons que les analyses portent sur quel type d'annotateur par exemple.

4.5 Résultats

4.5.1 Analyse de l'échantillon

Comme énoncé précédemment, nous avons voulu vérifier l'hétérogénéité de notre échantillon (A.5) à l'aide des données issues du Google Form que nous faisions passer à nos annotateurs à la suite des 4 annotations. Notre échantillon se présente de la façon suivante :

La répartition globale concernant la genre des annotateurs est quasiment égale, 52,2% sont des femmes et 47,8% sont des hommes.

2/3 de nos annotateurs ont entre 18 et 25 ans (62,2%), et 1/3 ont plus de 25 ans (26,7% ont entre 25 et 40 ans et 11,1% ont 40 ans et plus)

Un peu plus de la moitié des annotateurs sont étudiants (54,3%), les autres sont soit sans emploi (8,7%) soit salariés (37%).

La majeure partie de nos annotateurs est diplômé entre bac+1 et bac+5 (40% ont en dessous de bac+3, 37,5% ont entre bac+3 et bac+5). 17,5% sont en formation professionnelle et 5% ont un diplôme supérieur à bac+5.

Les domaines d'études sont aussi très variés, nous les avons synthétisé en nuage de mot, les mots en caractères plus importants sont les domaines les plus représentés dans notre échantillon :

FIGURE 12 – Domaines d'études des annotateurs

En général la mission était bien, voire très bien comprise, puisque seulement 15,2% au total ont indiqué avoir trouvé la mission difficile. Près de 3/4 des annotateurs ont trouvé que la durée était satisfaisante ou très satisfaisante, le reste l'ayant trouvé soit trop longue, soit moyennement longue (ils sont restés neutres).

Seulement 8,7% des annotateurs ont trouvé la plate-forme Glozz difficile ou moyennement compréhensible, le reste n'ayant pas eu de difficultés (56,5% l'ont trouvé compréhensible et 34,8% n'ont eu aucun problèmes).

Ce sont donc de bons retours que nous avons eu sur cette campagne, puisqu'en imaginant que nous devions enlever les annotations de personnes ayant eu des difficultés à comprendre la mission, une quarantaine d'annotations restent exploitables.

4.5.2 Analyse quantitative des relations et thèmes donnés

POINT DE VUE DES TEXTES

Avant toute analyse, nous vérifions si les arbres générés ont toutes leurs unités reliées par au moins une relation. Si tel est le cas, le nombre de relations est égal au nombre d'unités du texte - 1. Nous vérifions donc ceci pour chaque texte, à l'aide d'un algorithme qui renvoie vrai ou faux selon si le texte est entièrement relié ou non. Dans notre cas, l'entièreté des textes est totalement reliée, ce qui était attendu puisque nous précisions bien à l'oral, lors du texte "Bac-à-sable", que toutes les relations devaient être reliées. Nos annotateurs ont donc tous pris en compte cette consigne puisqu'aucune annotation possède d'unité sans relation.

Pour chaque texte, nous analysons la fréquence d'apparition des relations en fonction du nombre de relations totales. Le nombre de thèmes est aussi calculé relativement au nombre d'unités du texte, de façon à ce que la comparaison entre les textes de longueur différente soit

pertinente. Ainsi nos descriptions ci-dessous sont représentées sous forme de pourcentages, prenant en compte pour chaque analyse la longueur du texte associé.

Les thèmes

En moyenne, sur l'ensemble des 4 textes différents, le nombre de thème distincts donnés lors d'une annotation représente 16,03% du nombre d'unités totales. La répartition en fonction des textes est la suivante :

- bac à sable : 28,4%
- provocation : 7,7%
- nord : 14,9%
- florence : 13,3%

Pour provocation, le nombre de thèmes distinct représente seulement 7,7% des thèmes en moyenne, laissant alors suggérer que c'est le texte avec le moins de thèmes différents. Cela ne veut pas dire pour autant que peu de changements de thèmes sont présents, puisque ce chiffre n'indique en rien les alternances au niveau des thèmes.

Nous nous intéressons donc ensuite aux retours en arrière présents dans les textes. Un retour en arrière signifie que l'un des interlocuteurs revient sur une thématique précédemment discutée, alors que la thématique de la discussion avait changé.

- Dans Bac à sable, seulement 4,4% des annotations contiennent des retours en arrière.
- Seulement 2% des annotations du texte provocation contiennent des retours en arrière.
- Dans Nord, 8,7% des annotations contiennent des retours en arrière, dont un texte qui en contient 3 plus précisément.
- Dans Florence, 0,46% des annotations contiennent un retour en arrière, soit un seul texte.

Ces pourcentages très faibles ne nous permettent pas de repérer des changements de thèmes dans les textes donnés aux annotateurs. Sur Nord, le résultat proche de 10% mène quand même à réflexion. Pour tenter de vérifier si ces 8,7% peuvent être significatifs, nous pouvons vérifier si les annotateurs ayant trouvé ces retours en arrière dans Nord avaient tendance à en trouver dans les autres extraits. Si c'est le cas, cela mènera à penser que leur façon d'annoter menait à ce genre de résultats, et que ce résultat n'est donc pas significatif car biaisé par la manière d'annoter. Sinon, le résultat serait potentiellement exploitable mais il serait nécessaire d'avoir un plus grand échantillon pour vérifier la significativité du résultat.

Nous extrayons donc les 4 annotateurs ayant trouvé des retours dans Nord, et vérifions s'ils en ont trouvé dans d'autres extraits. Ces annotateurs sont A15, A01, B03 et B07. Seul B03 a aussi trouvé un retour en arrière dans provocation, les autres n'en ayant trouvé dans aucun autre extrait. Il semble donc que les retours en arrière dans Nord ne sont pas inférés par une manière d'annoter, et il serait intéressant d'avoir un échantillon plus grand pour vérifier la significativité de ces retours.

Les relations

Sur l'ensemble des 4 textes différents, ce sont les élaborations descriptives (28,9% des relations en moyenne) et les narrations (20,66%) qui sont utilisés en majorité. En affichant la moyenne de la fréquence d'apparition des relations en fonction de chaque texte, nous remarquons que quel que soit le texte, les relations les plus utilisées sont les "Narrations" et les "Élaborations descriptives".

Fréquence d'apparition des relations: bac à sable

	nb meta	nb phatique	nb elab evaluative	nb contreelab	nb narr	nb Elaboration descriptive	nb Conduite	nb Question	nbRéponse	nb Elaboration prescriptive
mean	1.269841	13.571429	9.828042	0	16.084656	17.301587	13.902116	13.280423	12.539683	0
std	4.111415	9.116528	7.596971	0	7.737563	10.006536	9.768531	6.514109	6.783673	0
min	0.000000	0.000000	0.000000	0	0.000000	0.000000	0.000000	0.000000	0.000000	0
max	14.285714	28.571429	33.333333	0	42.857143	42.857143	28.571429	33.333333	28.571429	0

Figure 13 – Fréquence d'apparition des relations - Bac-à-sable

Fréquence d'apparition des relations: florence

	nb meta	nb phatique	nb elab evaluative	nb contreelab	nb narr	nb Elaboration descriptive	nb Conduite	nb Question	nbRéponse	nb Elaboration prescriptive
mean	2.318841	6.932633	5.463582	0.668896	17.001380	22.775522	7.236290	18.458221	17.466821	1.677815
std	4.524899	6.865482	5.761005	2.191422	9.657107	11.074620	5.913316	6.903414	5.453764	3.672512
min	0.000000	0.000000	0.000000	0.000000	6.666667	0.000000	0.000000	7.692308	7.692308	0.000000
max	15.384615	23.076923	23.076923	7.692308	53.846154	38.461538	23.076923	38.461538	38.461538	16.666667

Figure 14 – Fréquence d'apparition des relations - Les deux Florence

Fréquence d'apparition des relations: nord

	nb meta	nb phatique	nb elab evaluative	nb contreelab	nb narr	nb Elaboration descriptive	nb Conduite	nb Question	nbRéponse	nb Elaboration prescriptive
mean	0	19.809973	6.005804	0.543478	28.52342	38.265236	4.678176	0	1.630435	0.543478
std	0	9.480325	6.980352	2.722888	17.18189	16.941696	8.468850	0	6.483494	2.722888
min	0	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0	0.000000	0.000000
max	0	33.333333	25.000000	16.666667	75.00000	72.727273	25.000000	0	33.333333	16.666667

Figure 15 – Fréquence d'apparition des relations - Nord

Fréquence d'apparition des relations: provocation

	nb meta	nb phatique	nb elab evaluative	nb contreelab	nb narr	nb Elaboration descriptive	nb Conduite	nb Question	nbRéponse	nb Elaboration prescriptive
mean	0.692109	14.316590	7.493971	1.671659	20.961844	37.026416	8.161557	3.303750	5.190631	1.181474
std	1.652274	6.682134	6.721273	3.362186	10.882723	13.102707	6.795102	2.576397	7.484069	2.607394
min	0.000000	4.545455	0.000000	0.000000	4.545455	8.695652	0.000000	0.000000	0.000000	0.000000
max	4.761905	31.818182	27.272727	13.636364	45.454545	59.259259	28.000000	9.090909	36.363636	9.090909

FIGURE 16 – Fréquence d'apparition des relations - Provocation

POINT DE VUE DES ANNOTATEURS

Globalement, les annotateurs ont tendance à mettre beaucoup de narration et d'évaluations descriptives, cependant, il est normal que ce pourcentage soit plus élevé car il en est de même dans les annotations de nos annotateurs experts.

nb Méta-question	1.063296
nb Phatique	13.615464
nb Élaboration évaluative	7.158372
nb Contre Élaboration	0.721008
nb Narration	20.695079
nb Élaboration descriptive	28.936393
nb Conduite	8.502908
nb Question	8.738463
nb Réponse	9.174847
nb Élaboration prescriptive	0.850692

TABLE 3 – Pourcentage moyen de relations sur l'ensemble des annotateurs

4.5.3 Analyse des arbres

Le premier point à vérifier du point de vue des arbres, est la structure globale de ceux-ci. En effet, si l'arbre issu de l'annotation génère deux sous-arbres, cela correspondrait à une remontée sans consistance dans un nœud ultérieur, ce qui correspond à une rupture conversationnelle.

Le premier algorithme que nous avons décidé de mettre en place est donc un algorithme détecteur de sous arbre. Il nous permet de savoir pour chaque texte, si plusieurs sous-arbres sont générés par l'annotation. Nous représentons donc ci-dessous pour chaque texte, le pourcentage d'annotations ayant généré des sous-arbres, et celui ayant généré des arbres bien formés.

FIGURE 17 – Sous Arbre dans les annotations de "Bac-à-sable" et "Provocation"

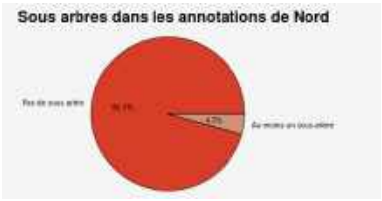


FIGURE 18 – Sous Arbre s les annotations de "Les deux Florence" et "Nord"

En regardant les graphiques générés ci-dessus, on s'aperçoit que seulement 4,3% des annotateurs génèrent des arbres mal formés dans Nord, et 7% dans le Bac à Sable. Cependant, une plus grosse proportion en génère dans Provocation (23,3%) et Florence (17,4%). En comparant ceci avec les annotations de nos annotateurs experts, nous remarquons une certaine corrélation puisque ceux-ci avaient effectivement généré des sous-arbres mal formé dans ces deux textes.Même si ces pourcentages ne sont pas significatifs, ils permettent quand même de mettre en avant le fait que près d'un quart des annotations génèrent le même type de résultats que ceux de nos annotateurs ré nts.

5 Conclusion

Le but de la mise en place de cette campagne était de trouver un protocole simple pour faire passer un maximum d'annotations à des annotateurs non-experts.

Grâce au travail d'anciens projets tutorés, nous avons pu gagner du temps sur la mise en place de celle-ci, en reprenant et améliorant ce qui avait déjà été fait. Ainsi, nous passions d'un guide complexe et long à un guide simplifié et très concis, et d'une procédure fastidieuse à une procédure simple et courte.

Ces changements dans la procédure entière de campagne nous ont permis d'obtenir plus de volontaires, de les maintenir concentrés durant toute l'annotation (celle-ci ne durant que 30 min en moyenne), et de leur faire annoter plus de textes que précédemment (4 textes annotés, dont 3 issus d'un corpus de conversations pathologiques).

Le déroulement a été un succès puisque 46 annotateurs se sont portés volontaires, chacun devant annoter 4 textes. Nous n'avons eu aucun abandon et l'ensemble des données obtenues est exploitable. En effet, nous remarquons après analyse que la consigne principale qui influe sur l'exploitabilité des données, a bien été comprise par l'ensemble de nos annotateurs : toutes les unités sont bien reliées par au moins une relation.

Nous avons aussi pu adapter et créer des outils de traitement de données et d'analyse des arbres à nos besoins, livrant un outil générique et réutilisable pour les futures campagnes d'annotation.

L'analyse des ruptures générées par la création de sous-arbres lors de l'annotation montre bien qu'il y a une similitude entre les textes ou nos référents annotateurs en avaient trouvé et ceux ou un plus grand nombre de nos annotateurs en ont trouvé.

Bien entendu, des améliorations sont encore possibles dans le but de pouvoir mener des campagnes toujours plus pertinentes. L'idée que nous avions eu au début de la campagne, qui est de développer un outil accessible en ligne pour faire passer l'annotation, est à considérer dans le futur. En effet, par manque de temps nous avions décidé de garder l'outil Glozz et de faire passer les annotations en présentiel, mais le fait d'avoir un outil en ligne permettrait d'augmenter grandement le nombre d'annotateurs. Effectivement, cela apporterait certainement un biais puisque nous n'aurons pas le contrôle de l'attention de l'annotateur, il pourrait donc faire ces annotations sans grande attention, en faisant autre chose à côté par exemple. Ce facteur est à prendre en compte et il serait intéressant de vérifier s'il est négligeable ou important sur un grand échantillon.

Bibliographie

[1] Joan Busquets, Laure Vieu, and Nicholas Asher. La sdrt : une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum*, 23(1) :73–101, 2001.

[2] Manuel Rebuschi, Maxime Amblard, and Michel Musiol. Schizophrénie, logicité et perspective en première personne. *L'Évolution Psychiatrique*, 78(1) :127–141, 2013.

A Annexe

A.1 Annexe 1 - Calendrier

Novembre à Janvier	Lecture de la bibliographie
Fevrier	Prise en main de Glozz + Rédaction du guide d'annotation
Première semaine de Mars	Améliorer le guide + Ajuster Glozz
Deuxième semaine de Mars	Précampagne : Recruter les annotateurs
20 au 31 Mars	Précampagne : Pré-annotations
3 au 7 Avril	Amélioration définitive du guide et de Glozz
8 au 23 Avril	Campagne : Annotations en dehors de Nancy
24 au 30 Avril	Campagne : Annotations à Nancy
2 premières semaines de Mai	Analyse des résultats + rédaction du rapport
2 dernières semaines de Mai	Finalisation du rapport

Table 4 – Calendrier prévisionnel

Guide d'annotation

La mission.

Votre mission, si vous l'acceptez, va consister en l'annotation d'un texte, illustrant une conversation entre une personne et un psychologue, vous ne connaitrez pas explicitement l'identité de chaque interlocuteur.

Dans quel projet de recherche suis-je impliqué ?

Le projet SLAM (Schizophrénie et Langage : Analyse et Modélisation) vise à modéliser des conversations entre un psychologue et une autre personne.

A quoi va servir mon annotation ?

Une annotation en linguistique consiste à ajouter de l'information à un texte/document.
Ces annotations vont permettre de cibler des particularités dans le discours.

Comment annoter ?

Dans notre cas, nous allons vous faire utiliser un logiciel, le logiciel Glozz. Celui-ci est installé et ouvert devant vous. Le texte présent est celui que vous allez annoter, ce texte est un entraînement, nous serons là pour vous accompagner afin que vous ayez une meilleure prise en main du logiciel, ensuite nous vous laisserons en autonomie sur trois autres textes.

Le but est de **relier des unités** (phrase ou segments de phrase) grâce à des **relations rhétoriques** (par exemple: réponse à une question, narration, etc), et de préciser les **thèmes** associés à chaque partie de texte (exemple : le locuteur parle de "la mort", de "sa vie", etc.).
Remarque : Notez que chaque unité est reliée à une unité qui la précède, mais pas forcément à l'unité qui la précède immédiatement.

Le tableau au dos du guide explique comment créer les relations pas-à-pas, et comment définir les thèmes, à l'aide de seulement quelques boutons.
Les unités sont déjà créées, elles sont surlignées sur le texte présent devant vous.
Chaque couleur d'unité correspond à un interlocuteur. De plus, pour plus de clarté, le changement d'unité est signalé par un passage à la ligne.

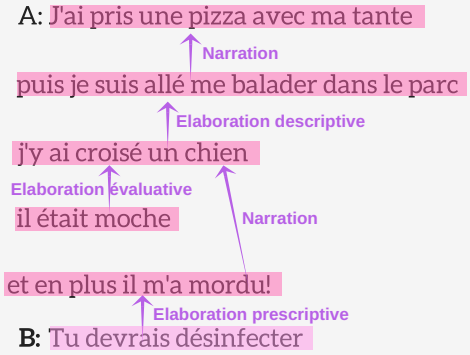
Veillez à bien suivre chaque étape et n'hésitez pas à nous appeler en cas de problème technique. Nous ne pourrons cependant pas vous guider sur le choix des relations.

Lorsque le texte est totalement annoté, appelez-nous pour que l'on enregistre votre résultat.

La fiche de synthèse ci-contre vous permet de comprendre et de visualiser les différentes relations rhétoriques possibles.

Narration
Suite d'éléments, qui permettent de raconter le déroulement d'une histoire

Elaboration descriptive
Phrase qui entre dans le détail et qui apporte des informations supplémentaires

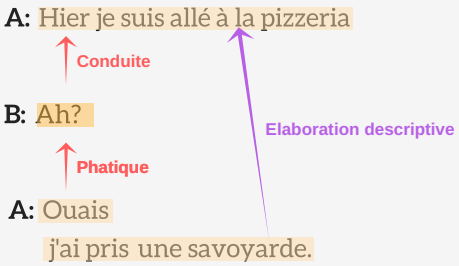


Elaboration prescriptive
Ordre ou conseil expliquant quoi faire ("il faut", "on doit", etc.)

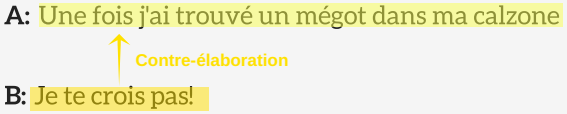
Elaboration évaluative
Donne une évaluation à la phrase ("bien", "super", "trop fort", "moche", "bon", etc.)

Conduite
Question/Intervention qui vise uniquement à relancer l'interlocuteur

Phatique
Prolonge la conversation, sans servir à communiquer un message



Contre-élaboration
Manifeste un désaccord, mène à une riposte

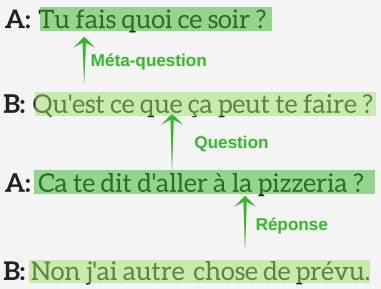


Question
Demande faite pour obtenir des informations. Une question se termine par un point d'interrogation

Réponse
Enoncé suite à une **question** posée uniquement

Méta-question
Question portant sur la conversation plutôt que sur le contenu de celle-ci

La flèche, représente pour vous le SENS de la mise en relation de 2 unités.



Utiliser Glozz

Relations	 <u>créer une relation</u> 1) cliquez sur l'icône correspondante 2) dans la liste des relations, sélectionnez celle qui vous semble la plus pertinente en respectant les règles de la fiche 3) relier les unités entre elles en respectant le sens des flèches indiqué dans la fiche **ATTENTION :** Avant de créer la première relation rhétorique, reliez la **première unité** du texte au **mot "Début"** à l'aide de la **relation "Narration".**	 <u>éditer ou supprimer une relation</u> • **modifier** en bougeant le début ou la fin de la relation vers une autre unité • **supprimer** en sélectionnant l'unité (un clic dessus) puis en appuyant sur la touche suppr
Thèmes	 <u>ajouter un commentaire</u> Chaque fois que vous remarquez un changement de thème dans la lecture, placez un drapeau au niveau de l'unité qui correspond au début du nouveau thème. Il faudra ensuite définir un mot clé pour le thème. <u>REMARQUES:</u> Utilisez **10 thèmes <u>maximum</u>** par texte. Lorsque vous utilisez un thème, reprenez-le mot pour mot si vous considérez qu'il revient.	

A.3 Annexe 3 - Relations rhétoriques

Relations	Explications	Exemples
Narration	Suite d'éléments qui permettent de raconter le déroulement d'une histoire	P1 : "Guy a pris un repas fantastique. Il a mangé du saumon. Il a dévoré plein de fromage.
Elaboration Descriptive	Phrase qui entre dans le détail et qui apporte des informations supplémentaires. L'élaboration descriptive se traduit par une sorte de "zoom", on part d'un fait et on rentre dans les détails.	P1 : "Je ne suis pas de bonne humeur, il pleut et je suis mouillée."
Elaboration évalutative	Donne une évaluation à la phrase ("bien", "super", "moche", "beau", "bon", etc.)	P1 : Michel Sardou a sorti une nouvelle chanson. P2 : Elle est trop bien !
Elaboration prescriptive	Ordre ou conseil expliquant quoi faire ("il faut" , "on doit")	P1 : J'ai super mal au ventre. P2 : Tu devrais prendre un médicament.
Contre-élaboration	Manifeste un désaccord, mène à une riposte.	P1 : "Je parle cinq langues différentes" P2 : "C'est impossible !"
Conduite	Question/Intervention qui vise uniquement à relancer l'interlocuteur	P1 : "Hier je suis sortie en boite !" P2 : "Ah ?"
Phatique	Prolonge la conversation sans servir à communiquer un message	P1 : Je passe te chercher demain après le boulot. P2 : D'accord
Méta-question	Question portant sur la conversation plutôt que sur le contenu de celle-ci	P1 : J'ai perdu tes clefs. P2 : Hein tu peux répéter ?
Question	Demande faite pour obtenir des informations. Une question se termine par un point d'interrogation.	P1 : Hier je suis allé au cinéma P2 : Tu es allé voir quoi ?
Réponse	Énoncé suite à une question posée uniquement	P1 : Tu as mangé quoi hier ? P2 : Un bon couscous.

A.4 Annexe 4 - Textes

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 19 – Texte "Les deux Florence"

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 20 – Texte "Provocation"

Debut

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,

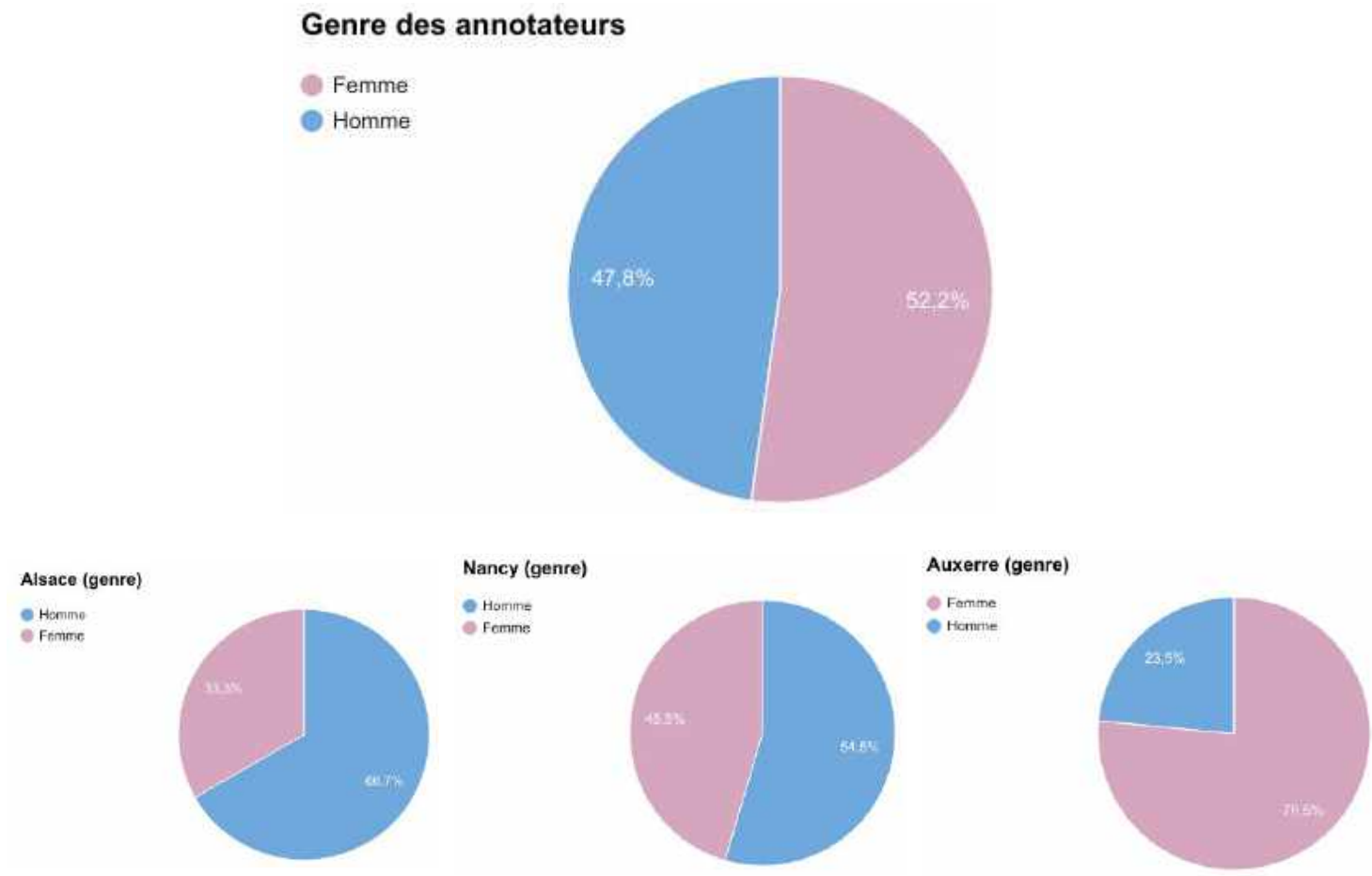
c'était vraiment super bon.

FIGURE 21 – Texte "Bac-à-sable"

A.5 Annexe 5 - Échantillon

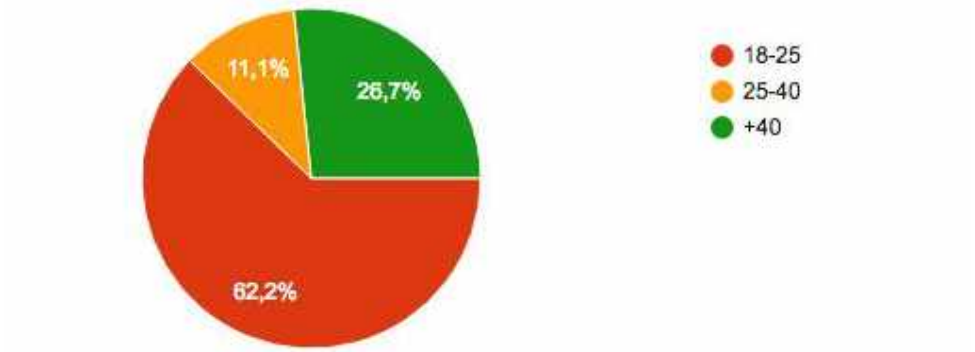
Nous avons eu un total de 46 annotations : 18 annotations en Alsace, 17 annotations à Auxerre, 11 annotations sur Nancy. Pour chaque information sur l'annotateur, nous faisons une présentation globale ainsi qu'une présentation par lieu.

Genre:

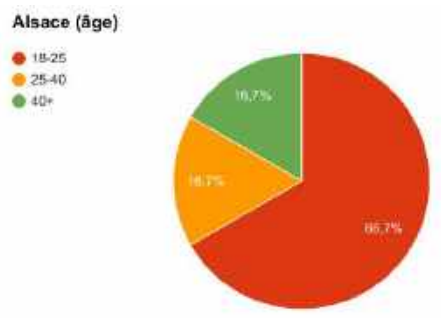


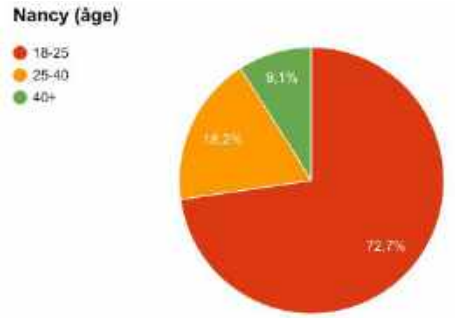
Tranche d'âge:

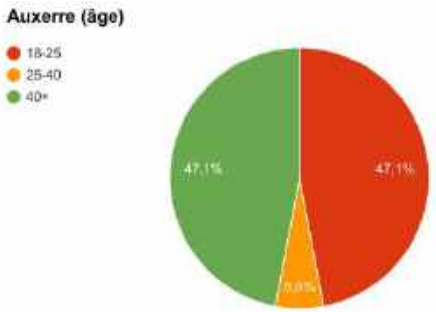
Nous avons répartis les tranches d'âges de la façon suivante : 18-25, 25-40, 40 et plus.



⅔ de nos annotateurs ont entre 18 et 25 ans (62,2%), ⅓ ont entre 25 et 40 ans (26,7%) et 40 ans et plus (11,1%).

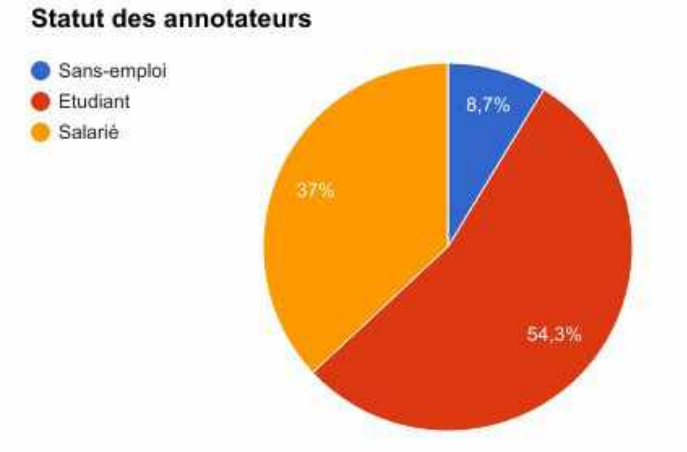


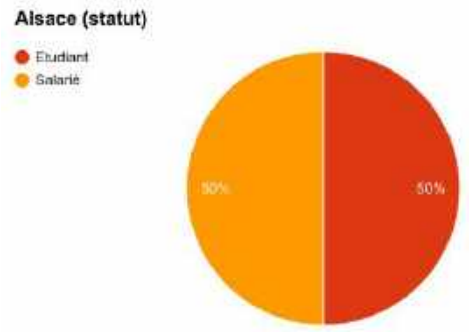


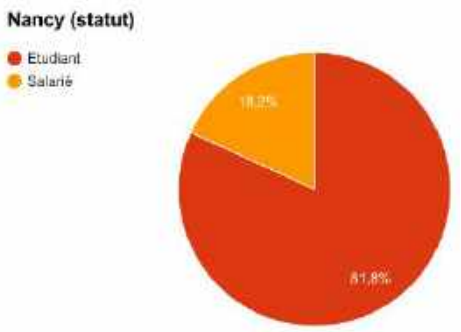


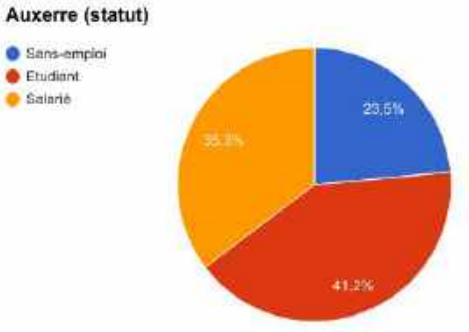
<u>Statut social:</u>

Concernant le statut des annotateurs nous avons la répartition générale suivante :

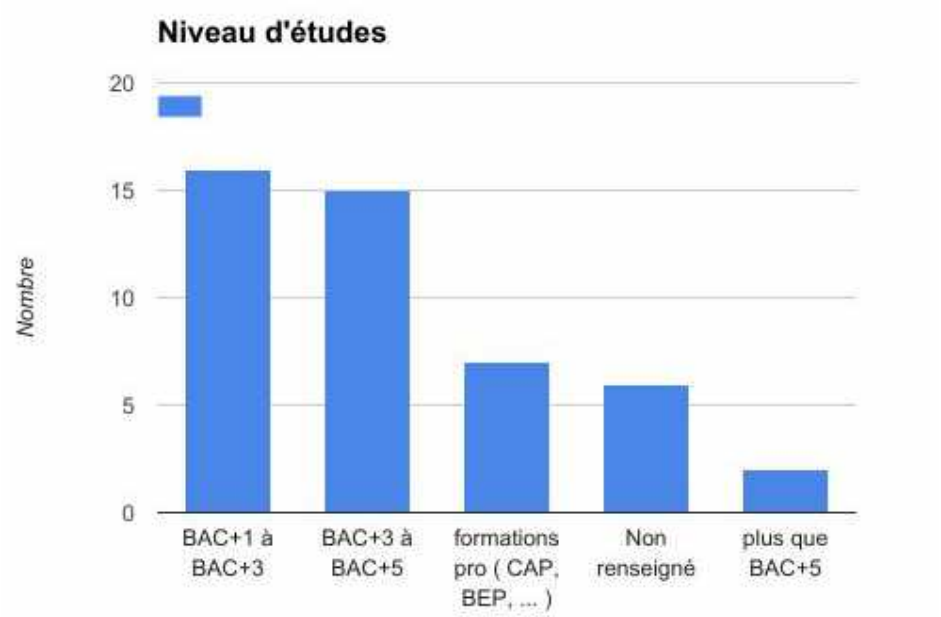


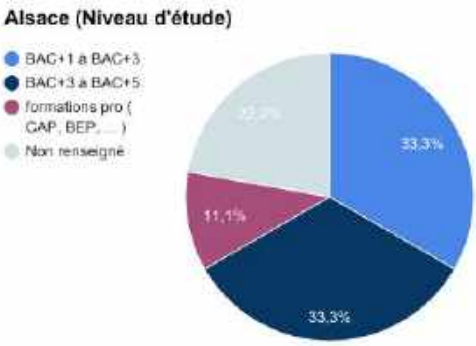


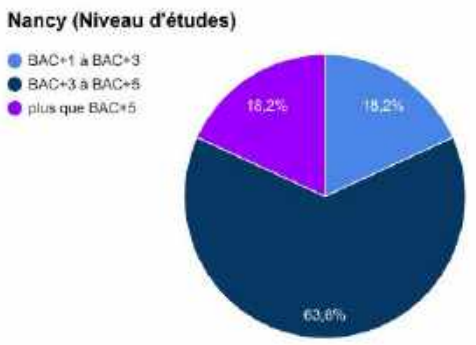


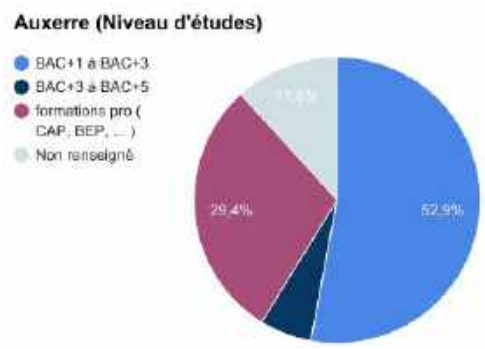


Niveau d'études:









Domaine d'études :

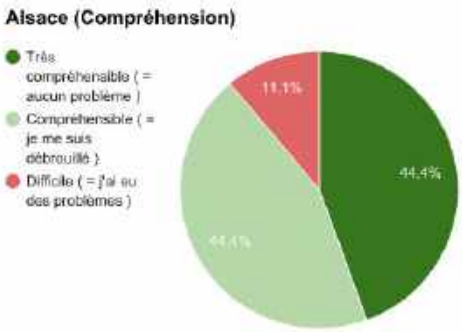
En Alsace nous ressassons les domaines d'études suivants : Biologie, Physique, Informatique (3), Psychologie du travail, Chimie, Carrières juridiques, Banque, Sage-Femme, Institutrice, Mécanicien.

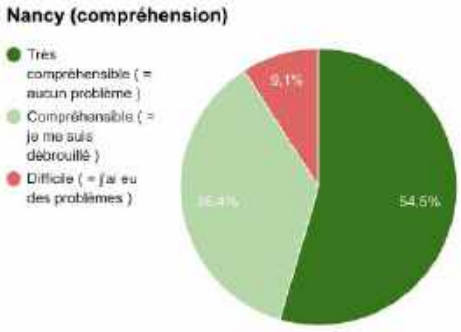
A Nancy nous ressassons les domaines d'études suivants : Sciences cognitives (5), MIAGE (Méthodes Informatiques Appliquées à la Gestion des Entreprises) (2), TAL (Traitement automatique des langues), Linguistique, Informatique

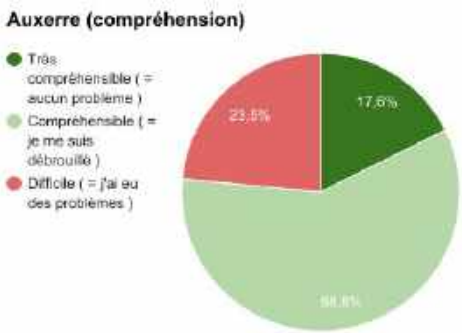
A Auxerre nous ressassons les domaines d'études suivants : Notariat, Psychologie, Institutrice, Environnement, Sciences/Mathématique, Droit (2), Infirmière, Communication d'entreprise, Orthophonie.

L'ANNOTATION:

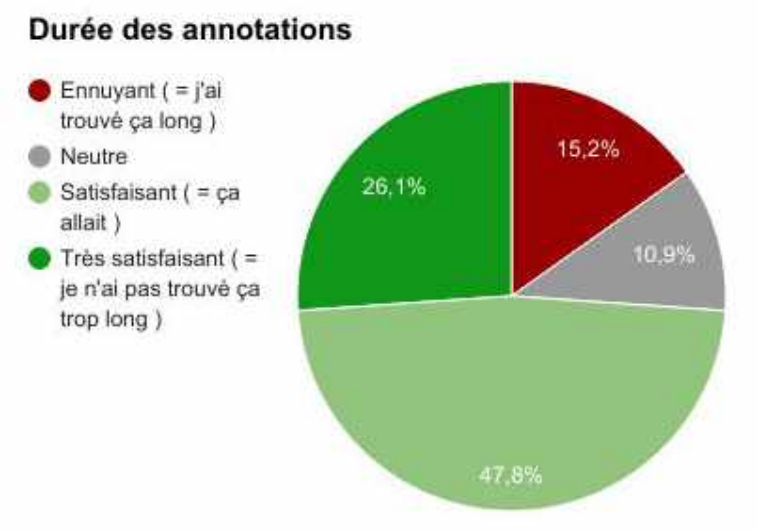
La compréhension de la mission globale:

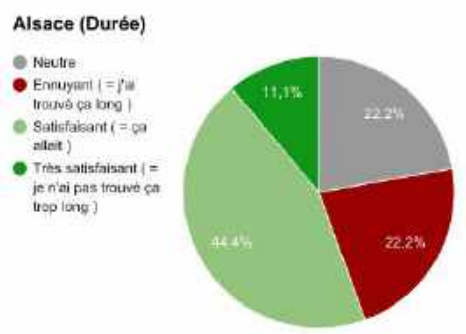


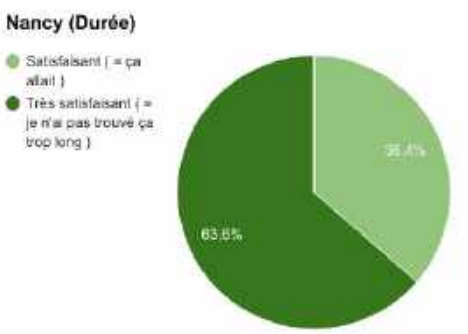


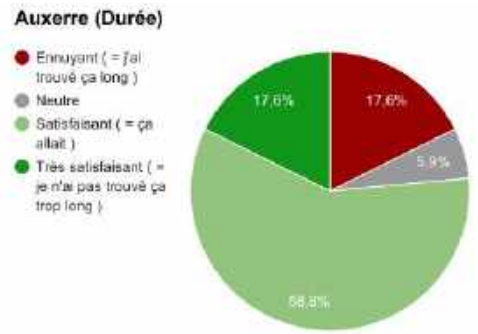


Durée d'annotation:

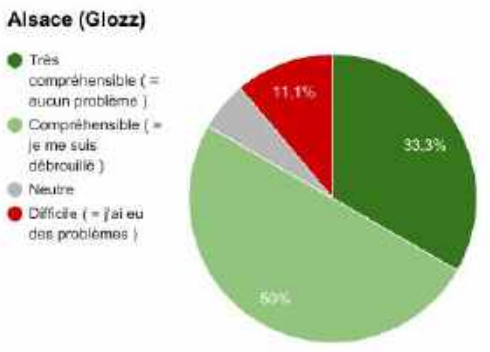


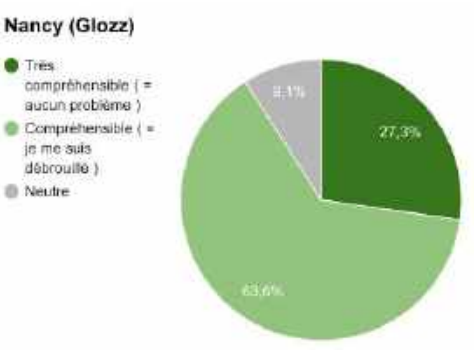


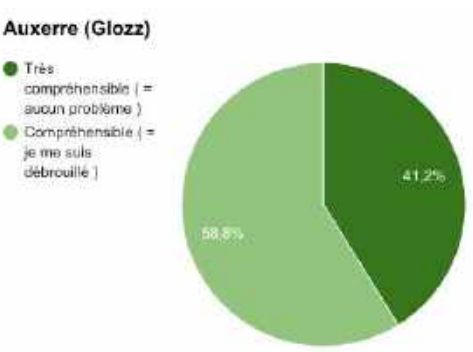




Utilisation de Glozz :


Alsace (Glozz)
Très compréhensible (= aucun problème)
Compréhensible (= je me suis débrouillé)
Neutre
Difficile (= j'ai eu des problèmes)
11,1%
33,3%
55%


Nancy (Glozz)
Très compréhensible (= aucun problème)
Compréhensible (= je me suis débrouillé)
Neutre
9,1%
27,3%
63,6%


Auxerre (Glozz)
Très compréhensible (= aucun problème)
Compréhensible (= je me suis débrouillé)
41,2%
58,8%