“ Qu’est ce qu’il dit ? ”: analyse de marqueurs audiovisuels de l’incompréhension

Éric Le Ferrand

LE FERRAND
Éric

Sous la direction de Maeva Garnier et Fabien Ringeval

Laboratoire : GIPSA-lab et LIG

UFR LLASIC
Département Sciences du Langage et FLE

Mémoire de master 2 Science du Langage - orientation Recherche - 20 crédits

Parcours : Industrie de la Langue

Année universitaire 2017-2018

"Qu'est ce qu'il dit?" Analyse de marqueurs audiovisuels de l'incompréhension

LE FERRAND
Éric

Sous la direction de Maëva Garnier et Fabien Ringeval

Laboratoire : GIPSA-lab et LIG

UFR LLASIC
Département Sciences du Langage

Mémoire de master 2 Sciences du Langage orientation Recherche - 20 crédits

Parcours : Industrie de la Langue

Année universitaire 2017-2018

Remerciements

Je tiens à tout d'abord à adresser mes remerciements les plus sincères à mes directeurs de mémoire, Maëva Garnier et Fabien Ringeval, qui m'ont accompagné, aidé et soutenu durant toute la durée de ce projet. Ils ont été encourageants et m'ont aidé à progresser.

Je remercie aussi Xavier Laval qui a créé notre bouton poussoir pour pouvoir mener à bien notre expérience, Frédéric Elisei qui nous a accompagné lors de nos enregistrements et Silvain Gerber qui nous a accompagné pour nos analyses statistiques.

Je remercie également ma mère et Éva Barloy qui ont pris de leurs temps pour la relecture de ce mémoire.

Et enfin je remercie mes amis et collègues William, Mahault, Céline, Nina, Alexia et Antonio, ainsi que mon père, ma sœur et mon frère pour leur soutien et encouragements.

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : LE FERRAND

PRENOM : ÉRIC

DATE : 31/08/2018

Sommaire

Introduction

Cette recherche a pour but de caractériser des marqueurs d'incompréhension liée à la perturbation du canal de communication dans le dialogue face à face. Ce mémoire fait suite à un stage de recherche effectué dans le cadre du Master 2 en sciences du langage, parcours industrie de la langue à l'Université Grenoble Alpes. Ce travail de recherche a été encadré par Maëva Garnier, chercheuse CNRS dans l'équipe PCMD (Perception, Contrôle, Multimodalité et Dynamique de la Parole) du GIPSA-lab, et Fabien Ringeval, enseignant chercheur dans l'équipe GETALP (Groupe d'Étude en Traduction Automatique/Traitement Automatisé des Langues et de la Parole) du Laboratoire d'Informatique de Grenoble (LIG). Ce stage s'est déroulé dans ces deux laboratoires afin d'approfondir à la fois l'aspect cognitif et l'aspect computationnel de ce projet.

Nous ne sommes pas inactifs lorsque nous écoutons quelqu'un parler mais nous émettons un grand nombre de signaux. Ces signaux peuvent être le témoignage d'un état émotionnel ou alors l'expression d'un degré d'accord par rapport à un discours donné. Ce genre de signaux peut servir à gérer les tours de parole ou bien témoigner de la bonne compréhension d'un discours. Un grand nombre d'études ont déjà attesté la présence d'une communication non verbale mais on remarque que beaucoup d'entre elles ne prennent en considération que l'émetteur d'un discours et non son récepteur. Dans les quelques études qui prennent en compte ce dernier, certaines réactions ont été étudiées mais ça n'a pas été le cas des réactions portant sur l'incompréhension du discours.

Le but de notre recherche est donc de savoir s'il existe des comportements types reliés à incompréhension et comment ils se manifestent. Pour cela nous avons mis en place un protocole expérimental dans lequel nous avons mis des sujets en situation d'incompréhension. Nous avons ensuite annoté notre corpus créé avec les comportements de sujets puis nous avons procédé à un ensemble d'analyse statistique afin de savoir si nos comportements variaient selon nos conditions expérimentales. Nous avons finalement extrait un ensemble de paramètres depuis notre corpus sur lesquels nous avons fait un ensemble de tests afin de repérer parmi ces données celles qui pouvaient être exploitées pour un système de détection automatique.

Partie 1 : État de l'art

Chapitre 1. La communication interpersonnelle

La communication consiste en l'échange d'informations entre une source, que nous appellerons émetteur et un destinataire que nous appellerons récepteur. Cette communication se fait par le biais d'un canal qui peut être bruité ou non. Cet échange est ensuite bouclé par une rétroaction du récepteur vers l'émetteur. Un schéma de communication qui a été proposé par Shannon et collègues (1948) reprend ces idées avec tout de même des limites du fait que les différents actants de ce modèle sont considérés comme des entités abstraites et non comme des individus. Abric (1996) rappelle que les individus sont soumis à *"des facteurs psychologiques, des contraintes sociales, des systèmes de normes, des valeurs"*. On préférera donc sa définition qui parle d'un *"ensemble de processus par lesquels s'effectuent les échanges d'informations et de significations entre des personnes dans une situation donnée"*.

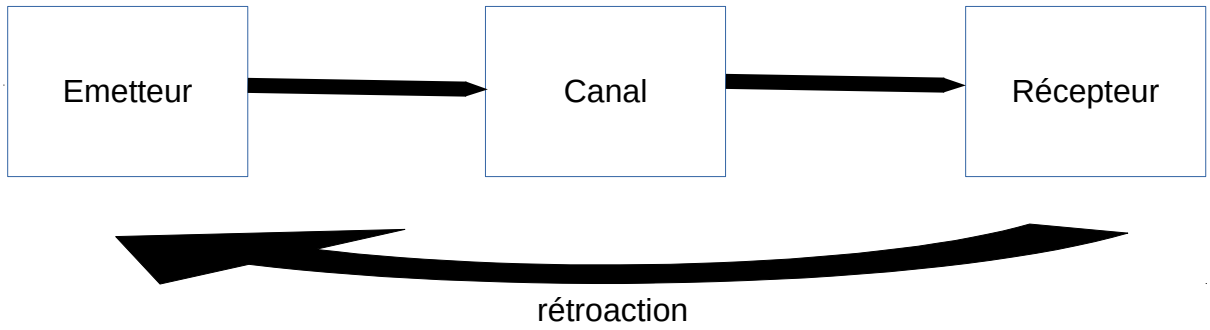


Figure 1: Schéma de communication selon Shannon

Les études sur la communication ne sont pas rares, cependant, la majorité d'entre elles se focalise sur l'émetteur du message et non sur le récepteur qui est loin d'être inactif lors de ce processus communicatif. Dans notre étude, nous allons nous centrer sur cette rétroaction produite par le récepteur et plus précisément sur les rétroactions qui témoignent de l'incompréhension d'un discours. Comment le récepteur la signifie-t-il face à l'émetteur ? Comment l'émetteur interprète-t-il ces signaux et comment s'en sert-il pour réguler son discours ?

Dans cette optique, nous allons d'abord aborder la construction du dialogue en évoquant le rôle des actants impliqués. Par la suite, nous allons nous intéresser au phénomène de rétroaction et les formes qu'il peut prendre, nous mettrons ensuite en avant ces rétroactions qui témoignent de l'intelligibilité d'un discours. Dans une seconde partie,

nous verrons les traitements qu'il est possible d'effectuer sur ces marqueurs, les mesures et leurs classifications.

1. Le Dialogue

Le dialogue est une forme particulière d'interaction. Chovil (1989) a démontré que plus une interaction se rapproche d'une interaction face à face, plus il y a de signaux qui viennent ponctuer cette interaction. En effet, pendant le dialogue, un certain nombre de signaux sont produits pour signifier d'un état émotionnel, pour gérer les tours de parole ou bien pour faire comprendre que l'on comprend le discours qui nous est adressé. Ces signaux peuvent être aussi bien verbaux que non verbaux. Dans son étude sur la compréhension mutuelle, Bavelas et collègues (2018) ont divisé le dialogue en trois étapes :

- L'émetteur donne une nouvelle information au récepteur ;
- Le récepteur reçoit l'information en émettant un signal verbal ou non verbal ;
- L'émetteur va, en fonction du signal reçu, soit continuer son discours soit apporter de nouvelles informations.

De manière générale, la bonne compréhension d'un discours peut être validée par la réponse verbale ou non verbale du récepteur qui va, dans ce cas, réagir de manière adéquate au discours de l'émetteur à travers des questions ou en reprenant les informations du discours ou juste en hochant de la tête. Une étude menée par Eberhard et Nicholson (2010) s'intéresse à l'importance du regard. La coordination de la compréhension entre les différents actants serait étroitement liée au regard. Bavelas décrit ce phénomène avec le terme "calibrating" un court instant pendant lequel les regards des deux interlocuteurs se croisent ce qui va permettre au récepteur d'un discours d'émettre des signaux vers l'émetteur pour témoigner sa participation au dialogue.

Il faut bien distinguer l'écoute active et l'écoute passive. Lorsqu'un récepteur écoute un discours de manière active, nous avons dit qu'il émet un grand nombre de signaux qui viennent enrichir l'échange. Un récepteur passif n'émettra pas de tels signaux et le discours de l'émetteur en sera appauvri et confus (Bavelas et al. 2011). Dans cette partie nous allons donc nous intéresser à l'impact de ces signaux dans le dialogue ainsi que leurs sens, puis explorer les différents aspects de la rétroaction.

1.1. Communication verbale et non verbale

La communication verbale est définie par un ensemble de sons émis par des individus dans le but de communiquer entre eux. Linguistiquement parlant, cette communication est composée d'unités lexicales, qui contiennent les informations sémantiques d'un discours, et la prosodie qui regroupe plusieurs phénomènes : l'accentuation ou l'intonation, qui donneront des informations sur la catégorie grammaticale de la phrase, et la pragmatique comme l'injonction, l'interrogation ou l'exclamation ou encore le témoignage d'un état émotionnel. Elle aura aussi la fonction de focus, c'est à dire de mettre en avant un élément particulier dans une phrase.

En parallèle, une gestuelle est présente pour accompagner le dialogue. L'hypothèse de Clark et Gerrig (1990) supposait que la gestuelle était partie intégrante de l'interaction et donc l'absence de récepteur pouvant percevoir ces signaux n'empêche pas ces signaux de se produire. Özyürek (2017) sépare les gestes en deux catégories, d'un côté les emblèmes, c'est à dire des gestes symboliques pouvant remplacer le discours, de l'autre, des gestes moins conventionnels dont le sens dépend du discours. Cette première catégorie ayant des caractéristiques sémantiques explicites, son côté volontaire lors de sa production n'est pas à remettre en cause, cependant, il n'en est pas de même pour la deuxième catégorie. Dans une étude menée par Bavelas (2007), on peut voir qu'aussi bien par le biais d'un téléphone que face à face, les individus produisent des gestes quand ils parlent. Le récepteur n'étant pas en mesure de voir ces gestes, leurs production pourrait donc être utile pour l'émetteur lui-même. Toutefois dans leur étude, Munhall et collègues (2003) ont montré que la compréhension du discours à travers un canal bruité était plus importante lorsque que l'émetteur était visible que quand il ne l'était pas, cette gestuelle est donc aussi utile pour le récepteur.

Le processus qui amène à la production de ces gestes reste flou. McNeill (1992) propose une théorie selon laquelle la production de gestes est un processus spontané et automatique et que la forme de ces gestes reflète la vision mentale d'un événement (Window Architecture). De Ruiter (2007) enrichit cette théorie en soutenant l'idée que ce processus est déclenché, avec le discours, de manière consciente mais la production même de ces gestes serait inconsciente.

1.2. La rétroaction

De manière générale, la rétroaction est une réponse produite par le récepteur afin de réagir à l'énoncé de l'émetteur et participer à la construction du dialogue. Comme nous

l'avons déjà dit, la réponse du récepteur n'est pas forcément verbale. Dans sa théorie de l'*appraisal*, Scherer (1999) s'intéresse au mécanisme de prise de décision face à une situation. L'*appraisal* peut être défini comme un processus direct, immédiat, non réflexif, automatique, par lequel les événements de l'environnement sont jugés comme bons ou néfastes à partir de la perception qu'on en a. Ce mécanisme est donc une prise de décision suite à un événement. Dans notre cas, le discours correspond à l'événement et la rétroaction à la réaction du récepteur. Cette réaction peut se manifester à travers des états émotionnels visibles.

À cette théorie, il faut ajouter la théorie de l'esprit (theory of mind) qui est définie par Krych-Appelbaum et collègues (2007) comme la faculté de pouvoir prédire ou comprendre l'état mental de quelqu'un d'autre. Rizzolatti et collègues (1996) montrent que certains neurones (neurones miroirs) rentrent en activité de la même manière si un sujet effectue une action ou voit quelqu'un d'autre effectuer cette même action. Ces neurones rentrent en activité lorsqu'il existe une action intentionnelle visant à un but. Ces expériences montrent donc une correspondance entre la réalisation d'une action et la compréhension de l'action par autrui grâce à la simulation cognitive de cette même action. Il en va de même pour les émotions. La manifestation d'une émotion entraîne chez autrui la simulation cognitive de cette même émotion (l'empathie). La communication non-verbale serait perçue intuitivement. C'est à partir des signaux émis par le récepteur vers l'émetteur que celui-ci peut se représenter l'état d'esprit dans lequel est son interlocuteur. Quelle émotion a t-il ? Comprend il le discours ?

La rétroaction peut avoir plusieurs fonctions. Bavelas a isolé trois fonctions principales (Bavelas et al., 2018) :

- Les rétroactions qui manifestent des émotions ;
- Les rétroactions pour gérer les tours de parole (turn taking) ;
- Les rétroactions pour signifier sa compréhension.

Cependant une telle catégorisation peut être considérée comme réductrice. McCarthy (2003) a décrit le phénomène de smallTalk comme un ensemble de micro interventions du récepteur signifiant qu'il est toujours attentif à ce qui est dit.

Tout comme il a été mentionné ci-dessus. La rétroaction contient aussi ces deux modalités : orale et gestuelle. On peut diviser ce phénomène de rétroaction en deux catégories spécifiques :

- La rétroaction générique ;
- La rétroaction spécifique.

Lors de l'écoute active, c'est-à-dire lorsque le récepteur participe activement au dialogue en manifestant des réactions au discours, on va appeler rétroaction générique, les signaux non verbaux qui apparaissent de manière répétée sans lien avec des points précis du discours. On les trouvera sous la forme de hochements de tête ou de sons tels que "mmh", "oui". Ces signaux sont omniprésents dans le dialogue. Ils sont considérés comme des rétroactions "par défaut" (Kiessling et al., 1993). Elles sont là pour faire comprendre à l'émetteur que l'on est toujours attentif à son discours. En revanche, certains de ces signaux peuvent aussi montrer la volonté du récepteur de prendre la parole. Dans une recherche menée par Ward (2006) il est montré que ces phénomènes ont plusieurs fonctions. Ils ont une grande importance dans la gestion des tours de parole, l'adhésion à un discours ou la bonne compréhension du discours.

A l'inverse, Bavelas et collègues (2011) ont aussi noté qu'un certain nombre de signaux étaient spécifiques à certains points du discours. En plus de communiquer la bonne compréhension d'un discours, ils véhiculent des informations émotionnelles ou empathiques. Ces signaux se distinguent des rétroactions génériques par leur synchronie avec des points précis du discours. Un haussement de sourcils pour signifier son étonnement face à l'annonce d'une nouvelle par exemple.

2. L'incompréhension

Dans le flot de signaux non verbaux qui est produits lors de la rétroaction, les témoignages de la bonne perception du discours apparaissent également. Très peu d'étude ont été menée pour isoler des marqueurs spécifiques à l'incompréhension en tant que telle.

Garrod et Pickering (2004) parlent d'un phénomène d'alignement interactif. Il s'agit d'un alignement entre la base de connaissance entre un émetteur et un récepteur. On parle alors de désalignement quand cette harmonie est rompue. Ce désalignement peut avoir plusieurs causes :

- Une non compréhension, c'est à dire un manque au niveau des références communes. Si l'émetteur emploie un mot inconnu par le récepteur par exemple.

- Une incompréhension (Weigand, 1999), c'est à dire une information mal perçue de la part du récepteur. Si le canal est bruité par exemple et que l'information est inaudible par le récepteur.

Notre étude se concentrera sur cette seconde situation.

Nous avons expliqué ci-dessus, que les rétroactions peuvent être génériques ou spécifiques. Nous partons du principe que l'incompréhension, étant un état ponctuel, sera exprimée pas des rétroactions spécifiques. Elles seront synchronisées avec le discours de l'émetteur ou lors de pauses dans le discours.

Mais une autre hypothèse veut que deux types de signaux puissent être produits : soit des signaux adressés à l'émetteur qui apparaîtront lors du gaze window, soit des signaux pour soi.

Dans cette partie nous allons essayer d'isoler ces marqueurs qui témoignent de l'intelligibilité d'un discours. Dans une première partie nous nous intéresserons aux marqueurs oraux puis aux marqueurs gestuels

2.1. Marqueurs oraux

Comme il a été dit précédemment, le récepteur produit en permanence pendant le dialogue des sons du type "Mmh" ou "ouais". Ces signaux sont là pour signaler à l'émetteur que nous sommes toujours attentif à son discours et que nous l'invitons à continuer. Nous pouvons supposer que l'absence de ces signaux génériques peut témoigner un manque de compréhension. Nous allons surtout mettre en avant ce même genre d'indice avec comme caractère principal la synchronie avec le discours de l'émetteur.

Ward (2006) établit le fait que dans ces micro-expressions, les paramètres prosodiques sont plus porteurs de sens que les unités lexicales en elles-mêmes.

Des indices prosodiques peuvent être pris en compte lors de leur décryptage. Gardner (1998) s'est intéressé aux particularités prosodiques des mots "mm" "mh" et "yeah" en anglais en mettant en avant une tendance à y baisser l'intonation en signe d'accord par rapport à une information.

Une étude menée par Munshin et collègues (2000) remarque une structure régulière du dialogue d'un point de vue prosodique. De son côté, Shriberg et collègues(1998) listent des indices pertinents sur la construction même du dialogue d'un point de vue prosodique, à savoir la F0, la durée, la longueur des pauses et le débit. Ces indices étaient avant tout décrits du point de vue de l'émetteur mais ils restent des pistes possibles pour notre sujet.

2.2. Marqueurs visuels

Tout d'abord, Allwood et collègues (2010) se sont intéressés aux mouvements de tête lors du dialogue. Dans leur étude ils ont mis en avant le fait que ces mouvements apparaissaient principalement lors du feedback mais ont essayé de savoir ce qui les déclenchait. Ils notent qu'ils servent principalement à l'emphase du discours ou pour accentuer un feedback déjà existant (comme les SmallTalk décrit par Gardner). Avant cela, Cerrato et collègues (2003), dans une étude similaire, ont isolé différents types de mouvements de tête :

- *Nod* qui consiste en un mouvement de la tête vers l'avant ;
- *Jerk* qui consiste en un mouvement de la tête en arrière ;
- *Shake* qui consiste à un mouvement de tête de la droite vers la gauche et vice versa de manière simple ou répétée ;
- *Waggle* qui consiste en un mouvement de tête d'arrière en avant de droite à gauche ;
- *Side way turn*, un seul mouvement de tête vers la droite ou la gauche.

L'intérêt pour les expressions faciales remonte à 1972 quand Darwin et collègues se sont interrogés sur leur universalité (Darwin et al., 1972). Bavelas (1995) s'est intéressée à ces signaux faciaux produits lors du dialogue. Tout comme les signaux décrits par Allwood et collègues (2010), certains de ces signaux faciaux sont exprimés de manière synchrone avec le discours de l'émetteur. Elle note des exemples de mouvements de sourcils relatifs à des états émotionnels (de la surprise par exemple) mais aussi pour créer de l'emphase sur un mot en particulier. Goodwin (1986) a aussi démontré que ce genre de signaux pouvaient apparaître lors de courtes pauses entre les différentes phrases de l'émetteur.

Il est cependant possible de supposer que les expressions faciales relatives à la surprise telles que celles décrites par Ekman, peuvent signifier une incompréhension. Ortony et Patridge (1987) différencient la surprise liée à la réaction à un événement inattendu et la surprise liée à l'absence d'un événement attendue. Cette deuxième définition est aussi décrite de deux manières différentes à savoir d'un côté le conflit entre

l’attente d’un événement et son inconsistance, de l’autre, un conflit entre un événement et l’idée que l’on s’en faisait. Ce dernier conflit peut s’appliquer dans le cas d’une partie du discours mal perçue, incomprise. Cette expression de surprise se manifeste toujours selon Ekman par un haussement de sourcils, un écarquillement des yeux et une ouverture de la bouche ou un relâchement de la mâchoire.

Chapitre 2. Aspects méthodologiques

Dans cette partie nous allons nous intéresser aux mesures qu'il est possible de faire sur ces marqueurs. Notre étude part d'une approche multimodale. Nous allons donc récolter un certain nombre de données brutes audios et vidéos. La question est de savoir ce qu'il est possible d'extraire ce ces signaux. Nous allons dans un premier temps voir quelles mesures il est possible de faire puis décrire les outils que nous allons utiliser dans cette optique. Par la suite, nous allons décrire des méthodes d'analyse statistique pour faire ressortir les marqueurs pertinents et approcher des techniques d'apprentissage machine dans le but de pouvoir détecter automatiquement ces marqueurs.

1. Extraction des données

Il n'existe pas de corpus conçu pour mettre en avant ces signaux d'intelligibilité. En revanche un certain nombre de corpus interactionnels ont été créés. Deux approches ici nous intéressent, d'un côté la parole spontanée, c'est-à-dire des situations d'interaction plus ou moins naturelles sans consigne précise, de l'autre de l'interaction orientée tâche, c'est-à-dire des situations d'interaction contrôlée autour d'une tâche précise. Lors de nos traitements, nous n'allons pas exploiter des corpus déjà existants dans la mesure où ils ne correspondent pas tout à fait à nos objectifs, mais nous pouvons nous en inspirer dans la création d'une tâche expérimentale afin des données relatives à notre axe de recherche.

Parmi les corpus en rapport avec notre sujet, nous pouvons citer le Corpus of Interactional Data (CID) (Bertrand et al., 2008). Il s'agit d'un corpus audio-visuel d'interaction en face à face autour d'un thème donné. Ce thème pouvait être l'évocation soit de conflits professionnels, soit d'événements insolites qui leurs étaient arrivés. Ce corpus comporte 8 heures d'enregistrement en français annotées pour les domaines phonétique, prosodique, syntaxique, discursif et mimo-gestuel.

Pour ce qui est des corpus orientés tâche, le corpus HCRC (Carletta et al. 1996) apparaît comme une référence. Il s'agit d'un corpus de type MapTask. Il est composé de 18 heures d'enregistrement audio. La tâche consiste en une interaction entre un instructeur et un sujet. L'instructeur possède une carte sur laquelle se présentent différents items ainsi qu'un chemin tracé. Le sujet a la même carte sans le chemin, l'instructeur doit donc indiquer au sujet le chemin à suivre sur sa carte.

1.1. Mesures sur les signaux verbaux

Nous pouvons distinguer d'ores et déjà deux types de signaux verbaux :

- Des signaux linguistiques qui sont pertinents d'un point de vue lexical ;
- Des signaux dont la signification est portée par des variations acoustiques et/ou phonétiques.

Dans une étude visant à la caractérisation de quatre types de rires, Tanaka et Campbell (2011) utilisent les formants comme paramètres afin de construire un modèle HMM (*hidden Markov model*). Ces paramètres pourraient s'avérer pertinents pour la caractérisation de potentiels marqueurs acoustiques voisés. Dans une étude similaire, Ito et collègues (2005) utilisent des coefficients MFCC (*mel-frequency cepstrum*) et Delta-MFCC pour la caractérisation de rires non voisé.

1.2. Mesures sur les données visuels

Dans ses recherches, Ekman et collègues (1976) ont élaboré le FAC (*Facial Action Coding*), un système de classification des différents mouvements possibles visibles du visage, dans le but de pouvoir distinguer tous les comportements du visage. Grâce à un tel système, il est possible de distinguer différentes mimiques faciales par la combinaison de mouvements faciaux de base. Ils créèent ainsi un tableau qui correspond aux unités d'action possibles sur le visage. Ces expressions faciales sont très courtes. Elles durent en général entre 250ms et 5s. Un tel système pourrait être exploité pour la description de nos marqueurs en se focalisant sur les unités d'action (AUs) principales (utilisées notamment dans la description des expressions des sentiments par Ekman) :

Au_1	Remontée de la partie interne des sourcils
Au_2	Remontée de la partie externe des sourcils
Au_4	Abaissement et rapprochement des sourcils
Au_5	Ouverture entre la paupière supérieure et les sourcils
Au_6	Remontée des joues
Au_7	Tension de la paupière
Au_9	Plissement de la peau du nez vers le haut
Au_{10}	Remontée de la partie supérieure de la lèvre
Au_{12}	Étirement du coin des lèvres
Au_{14}	Plissement externe des lèvres (fossettes)
Au_{15}	Abaissement des coins externes des lèvres
Au_{17}	Élévation du menton

Au_{20}	Étirement externe des lèvres
Au_{23}	Tension refermante des lèvres
Au_{25}	Ouverture de la bouche et séparation légère des lèvres
Au_{26}	Ouverture de la mâchoire
Au_{28}	Succion interne des lèvres
Au_{45}	Clignement des yeux

Une recherche menée par Fasel et collègues (2003) décrivent de manière précise une manière de mesurer les signaux faciaux. Trois paramètres sont à prendre en compte lors de cette analyse :

- Leur position ;
- Leur intensité ;
- Leur dynamique.

Il n'y a pas d'intensité absolue lors de cette analyse mais relative par rapport au visage neutre d'un individu donné. On décrit trois étapes à la formation de ces signaux :

- Le déclenchement ;
- Le maintien ;
- Le relâchement.

Étant donné le manque d'étude sur les marqueurs faciaux chez le récepteur, notre démarche est d'explorer les marqueurs présents chez l'émetteur lorsqu'il est en situation d'interaction et d'effectuer le même genre d'analyse lors de nos expérimentations chez le récepteur.

Dans son étude sur les mouvements de tête lors du dialogue, Allwood (2010) a mis en avant l'intensité et l'amplitude de ces mouvements. Il a mesuré le décalage en millimètre de la tête par rapport à son axe horizontal et son axe vertical.

2. Analyses statistiques

Une fois les paramètres extraits, une analyse statistique est indispensable pour faire ressortir les signaux fréquents et pertinents pour notre étude. Ces mesures pourront nous

apporter des informations telles que les fréquences absolues et relatives de nos signaux à travers notre corpus et la cooccurrence de certains signaux par rapport à d'autres, comme du SmallTalk ou des mouvements faciaux qui se réaliseraient en même temps.

2.1. Calculs de fréquence

La fréquence absolue correspond au nombre d'occurrence d'une instance. La fréquence relative correspond à la fréquence d'une occurrence par rapport au nombre d'occurrences totales :

$$F = \frac{\text{Nombre d'occurences}}{\text{Nombre d'occurences totales}}$$

2.2. Calcul d'information mutuelle

Pour les cas de signaux parallèles (une expression faciale accompagnée d'un mot par exemple), il est possible d'utiliser le calcul d'information mutuelle spécifique (Pointwise mutual information). Il s'agit d'une mesure d'interaction permettant de prendre en compte des relations non linéaires, c'est-à-dire entre deux événements indépendants ou fortement corrélés. Si on note N le nombre total de couples, TP_1 le nombre total des couples ou l'événement 1 noté $E1$ est présent, TP_2 le nombre total de couples où l'événement 2 noté E2 est présent, et PP, le nombre de couples où E1 et E2 sont tous les deux présents, le rapport entre probabilité observée et probabilité théorique (supposant l'indépendance) se formule ainsi :

$$R = \frac{N \cdot PP}{TP_1 . TP_2}$$

Plus ce rapport atteint une valeur importante, plus l'hypothèse d'indépendance peut être rejetée

2.3. Modèle mixte

Un modèle mixte consiste en l'analyse de variance (ANOVA) entre différentes conditions (effets fixes) en prenant en compte des effets aléatoires. Pour chaque paramètre, on cherche le modèle le plus simple pour expliquer au mieux la variance entre différentes conditions en utilisant une approche descendante basée sur la minimisation du Critère d'Information Bayesien (BIC). Une valeur "p" est alors donnée pour estimer la significativité des effets testés. La notation conventionnelles s'écrit :

- $p>0.05$ NS, si l'effet testé n'est pas significatif

- $p<0.05$*, $p<0,01$**, $p<0.001$***, si l'effet testé est significatif

Un modèle mixte ne s'applique qu'aux données réelles comprises entre plus et moins l'infini. Une alternative à cette limite est la régression beta qui utilise le même principe en passant par une transformation des variables.

Partie 2 - Constitution d'un corpus audio-visuel

Chapitre 3. Récolte de données

1. Conditions expérimentales

Pour isoler des marqueurs d'incompréhension, il était nécessaire de constituer un corpus audio-visuel afin de pouvoir capturer des comportements selon ces deux modalités. Dans la création de notre protocole expérimental nous avons opté pour une tâche de type Maptask afin de forcer l'écoute active de nos sujets (Bavelas et al., 2011). Notre tâche consistait à demander à un ensemble de sujet de tracer un chemin à travers des items sur une carte à partir d'instructions perturbées. Nous avons créé un ensemble de cartes en suivant le modèle du corpus HCRC (Carletta et al., 1996). Ce modèle nous semblait pertinent par la multiplicité des chemins possibles. Pour avoir des énoncés comparables entre eux, ceux-ci ont été créés en suivant un même modèle :

- Un axiome formulé sous trois formes différentes : "tu passes", "tu vas", "tu dois aller". Chaque axiome est réparti de manière proportionnelle dans l'ensemble des énoncés ;
- Une direction : "à droite", "à gauche", "au-dessus", "au-dessous". Chaque direction est également répartie de manière équitable ;
- Un substantif féminin au singulier pouvant être visuellement représentatif : "de la tasse", "de la chaise", "de la flèche" afin d'être constant dans la forme de nos énoncés.

Des mots de transition ont été aléatoirement introduits dans ces énoncés afin d'ajouter de la fluidité à leur enchaînement ("ensuite", "après", "puis"). Une coupure audio et vidéo est ensuite posée à un point précis de chaque énoncé :

- Soit sur l'axiome (P0). « Tu passes » ou « Tu vas » par exemple, ce qui correspondrait à un niveau de perturbation faible du fait que le sujet a toujours à disposition les informations nécessaires pour se diriger sur la carte ;
- Soit sur la direction (P1). « à gauche » ou « au dessus » par exemple, ce qui correspondrait à un niveau de perturbation modéré du fait que, malgré l'information manquante, le sujet à toujours l'information relative à l'objet pour se diriger ;
- Soit sur l'objet (P2). « de la chaise » ou « de la poire » par exemple, ce qui correspondrait à un niveau de perturbation fort étant donné que, même si le sujet

dispose de l'information relative à la direction, il ne peut pas savoir vers où se diriger sans l'information relative à l'objet.

Les ensembles d'énoncés contiennent aussi des instructions non perturbées (SP) afin d'avoir des comportements types de la compréhension pour pouvoir comparer nos comportements relatifs à l'incompréhension.

Nous souhaitons de plus isoler deux catégories de marqueurs :

- Les signaux pour autrui ;
- Les signaux pour soi.

Pour cela, nous divisons notre protocole en deux parties :

- Une partie avec interaction dans laquelle le sujet est libre d'interagir avec l'expérimentateur. L'expérimentateur pourra donc redonner l'information au sujet s'il en exprime le besoin. Dans cette phase nous nous attendons à trouver des signaux pour autrui sans pour autant exclure la présence de signaux pour soi.
- Une partie sans interaction dans laquelle le sujet ne peut pas du tout interagir avec l'expérimentateur. Nous nous attendons à trouver des signaux pour soi en excluant cette fois tous signaux pour autrui.

Nous ajoutons à ces deux phases d'interaction une phase *test* constituée d'une seule carte de 12 énoncés perturbés selon le même modèle.

Nous souhaitons disposer d'une répartition égale de chacune des conditions en situation avec interaction :

- Tu passes au dessous de la robe (AISP)
- ~~Tu passes~~ à droite de la chaise (AIP0)
- Tu dois aller ~~à gauche~~ de la poire (AIP1)
- Tu vas au dessus ~~de la chaise~~ (AIP2)

et en situation sans interaction :

- Tu vas au dessous de la croix (SISP)
- ~~Tu vas~~ à gauche de la poire (SIP0)
- Tu passes ~~à droite~~ de la poule (SIP1)

- Tu dois aller au dessus ~~de la table~~ (SIP2)

À partir de la combinaison de nos trois axiomes avec nos quatre positions et nos quatre types de perturbation. Nous obtenons 48 énoncés différents répartis en quatre cartes de 12 énoncés chacune pour les deux situations d'interaction pour un total de 96 énoncés par sujet.

Du point de vue graphique, les cartes ont été constituées en fonction des énoncés créés. Les objets ont été sélectionnés afin d'être les plus explicites possible (cf. Figure 2). Ces objets ont été disposés de telle sorte que le chemin ne puisse ni se croiser ni faire demi-tour et que le chemin à emprunter soit le plus économique. Dans certains cas, des précisions ont été rajoutées aux énoncés pour lever des ambiguïtés ("tu vas ~~au dessus~~ de la poire *par dessous la hache*")

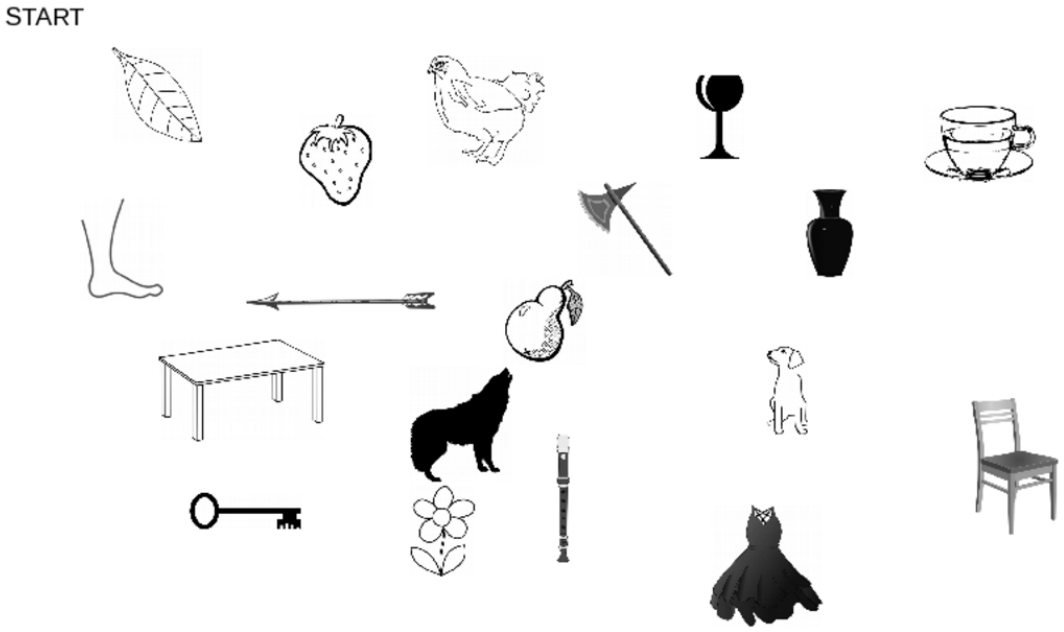


Figure 2: Exemple de carte

2. Protocole

Afin de mener à bien notre tâche, il était nécessaire de maîtriser le canal de communication. Pour cela, l'expérience s'est déroulée sous forme de vidéo conférence. Nous avons utilisé des prompteurs pour la diffusion vidéo afin d'éviter tout effet "faux-

jeton", (de Fornel, 1992) qui pourrait biaiser l'interaction, c'est à dire un décalage du regard causé par la position de la caméra dans les systèmes de vidéo-conférence. Le son est diffusé avec un casque audio et enregistré grâce à un micro casque placé à proximité des lèvres.

Le premier problème qui s'est posé concernait la forme de la perturbation à insérer dans l'interaction. Nous avons opté pour une coupure nette et instantanée du signal audio accompagnée d'un noir vidéo. Pour cela, un bouton poussoir a été confectionné et inséré dans l'installation. Du point de vue du son, nous disposions d'une carte son qui nous permettait de gérer les flux sonores. Le sujet recevait donc un son mixé contenant en partie la voix de l'expérimentateur et en partie sa propre voix pour simuler le réel. L'expérimentateur était placé dans une salle à part dans la même configuration que le sujet. Nous proposions d'abord une carte test afin d'habituer le sujet aux cartes puis, de manière alternée d'un sujet à l'autre, une phase avec interaction et une phase sans interaction. La situation d'interaction de la carte test était la même que la phase qui la suivait directement.

En situation d'interaction, l'expérimentateur se fiait aux retours du sujet dans l'enchaînement de ses énoncés. Le sujet était libre d'interagir avec l'expérimentateur qui redonnait, si besoin, l'information. En situation sans interaction, nous avions d'abord maintenu le modèle de la première phase avec un sujet pilote en lui précisant uniquement que la communication ne serait plus possible avec l'expérimentateur. Mais nous avons préféré par la suite substituer l'expérimentateur par une vidéo enregistrée au préalable afin de conserver les mêmes stimuli d'un sujet à l'autre. Le délai entre deux énoncés était de cinq secondes et le délai entre chaque carte était de douze secondes.

Nous avons donc appliqué ce protocole sur 14 sujets, 7 hommes et 7 femmes âgés de 20 à 60 ans (moy=28ans, ET=11,1). Tous les sujets ont le français comme langue maternelle, à l'exception du sujet n°15 dont la langue maternelle est l'arabe. Chaque sujet devait également remplir un questionnaire de personnalité issu du *big five personnality test* (Barrick et al., 1991). Cependant, les résultats de ces tests n'ont pas été exploités pour la suite de la recherche.

3. Biais et limites

Dans la situation sans interaction, parce qu'on ne corrige pas l'incompréhension au fur et à mesure, celle-ci reste constante pendant toute la phase. De l'incompréhension issus

d’énoncés précédents, se manifestent donc à la suite d’énoncés non perturbés, ce qui rend l'analyse des degrés d'incompréhension impossible pour cette phase.

Les impératifs techniques que nous avions, à savoir la maîtrise du canal de communication et la MapTask, plaçaient les sujets dans un cadre de communication inhabituel. Bien que nous ayons réussi à capter des signaux, il est possible de remettre en cause l'aspect écologique de cette expérience.

Bien que nous ayons alterné les conditions d'interaction d'un sujet à l'autre, nous avons proposé les mêmes cartes dans le même ordre pour chaque sujet. Il aurait été bénéfique de changer cet ordre afin de diminuer le risque de biaiser certaines parties du corpus.

Le fait que le français soit la seconde langue de l'un de nos sujets a posé un problème de compréhension dans les énoncés comprenant des items moins courants (la harpe et la tente). Les réactions liées à ces objets n'ont donc pas été prises en compte lors de notre analyse.

En outre, nous avons eu des problèmes techniques lors de l'enregistrement de certains sujets :

- L'enregistrement vidéo s'est interrompu pour le sujet n°2 ce qui nous a empêchés de récolter sa phase sans interaction
- L'enregistrement vidéo n’a pas démarré pour le sujet n°10
- Le sujet n°9 était mal positionné, ce qui a empêché la captation de son visage pendant la majeure partie de l'expérience

Ces trois sujets n’ont donc pas été pris en compte lors de nos analyses.

Chapitre 4. Organisation et annotation

1. Organisation

Une fois les données récoltées, il était nécessaire de les organiser afin de faciliter l'analyse.

Nous avons d'abord découpé nos vidéos afin de séparer la partie test (notées TAI ou TSI selon la condition d'interaction), la phase avec interaction (notée AI) et la partie sans interaction (notée SI). Le son enregistré par la caméra n'étant pas exploitable, nous avons dû resynchroniser le son, enregistré en parallèle, à la vidéo. Pour cela, nous avons appliqué une fonction d'inter-corrélation à chacun de nos fichiers pour récupérer le délai de décalage entre le son du micro et le son de la caméra, puis nous avons redimensionné et appliqué le son du micro aux vidéos correspondantes.

2. Annotation

À l'aide des outils ELAN et Praat, nous avons d'abord effectué une annotation générale sur trois sujets en repérant tous les comportements de l'interlocuteur qu'ils soient audibles ou visibles. Pour ce qui concerne les catégories audios, nous les avons discriminé en fonction :

- des parties de l'énoncé présentes dans la rétroaction (la direction et/ou l'objet) ;
- du destinataire de la rétroaction.

Pour ce qui concerne les catégories visuelles, nous avons étiqueté les catégories en fonction de

- la position du mouvement (les lèvres, la tête, les sourcils etc.) ;
- la forme du mouvement (étirement, contraction) ;
- plus rarement son caractère récursif, c'est à dire la répétition d'un même mouvement dans un court intervalle de temps, un hochement de tête par exemple. (en cas de caractère récursif isolé, nous avons associé le comportement à une catégorie existante en précisant cette spécificité).

Toutefois, poser un label sur un comportement n'a pas toujours été évident. Nous avons donc pour certains cas appelé ces catégories difficiles à étiqueter en fonction du lieu du mouvement et d'un numéro (divers tête n°7, par exemple). Après avoir isolé un

dictionnaire de 26 comportements, nous avons par la suite annoté les vidéos de tous les sujets en créant une ligne d'annotation par comportement, en enrichissant si besoin le dictionnaire. Les hochements verticaux n'ont pas été annotés car ils ont été catégorisés par des études précédentes comme marqueurs de compréhension il en est de même pour les rétroactions de type smallTalk présentes dans les conditions non perturbées qui sont aussi interprétées comme des signaux de compréhension. Par ailleurs, les mouvements incluant d'autres endroits que le visage et la tête n'ont pas été annotés à cause de la complexité du traitement que cela aurait engendré. Mis à part ces comportements très spécifiques, nous avons annoté tous les comportements visibles et audibles pour chacune des phases pour toutes les conditions.

En cas de comportement très spécifique à un sujet, nous avons créé des macros catégories dans lesquelles placer ces comportements. Toutefois, ce choix s'est montré problématique pour le sujet n°4, qui produisait un nombre important de comportements qui lui étaient propres. Nous avons donc créé une macro catégorie pour ce cas particulier.

En parallèle de ces comportements, nous avons annoté de manière automatique les coupures audio en réutilisant le script créé dans la réalisation de la vidéo diffusée pendant la phase sans interaction. Cette ligne d'annotation nous permettra de retrouver l'information relative à la latence d'une réaction par rapport à la perturbation. Nous avons aussi annoté les énoncés en indiquant le type de leur perturbation (SP, P0, P1 ou P2).

Toute l'annotation a été faite par un seul annotateur avec une relecture faite par ce même annotateur.

Partie 3 - Analyses

Chapitre 5. Analyses sur les marqueurs

Dans cette partie grâce à l'outil de statistique R, nous chercherons à savoir quel est le type des marqueurs observés. Nous ferons un tri de ces derniers pour ne retenir que ceux qui sont pertinents. Afin de procéder à cette analyse nous avons généré trois tableaux de données à partir de nos annotations :

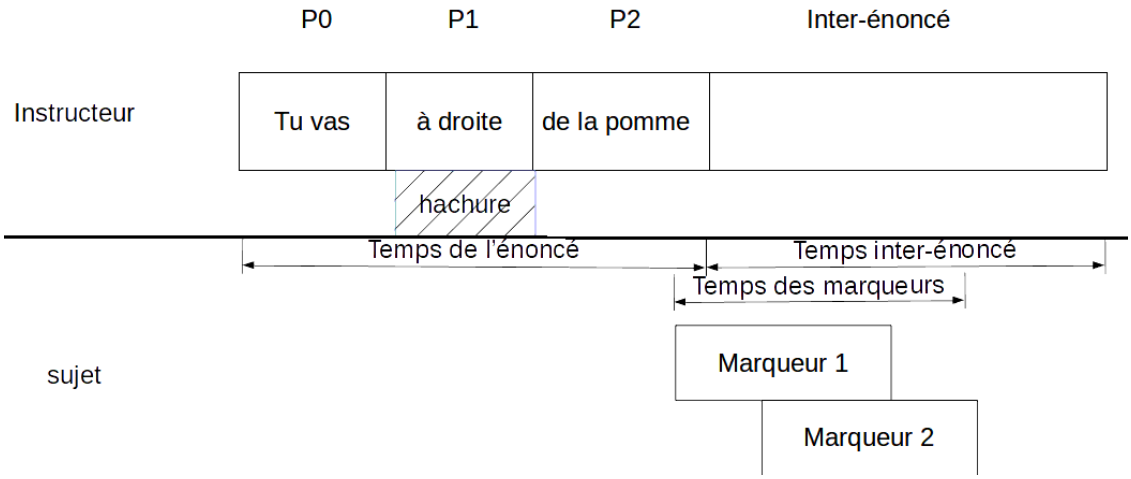


Figure 3: définition des intervalles

- Un tableau indiquant la fréquence des marqueurs en fonction des phases expérimentales, sur trois intervalles (cf. Figure 3) : le temps de l'énoncé, le temps inter-énoncé, le temps des marqueurs
- Un tableau indiquant pour chaque marqueur sa durée et ses latences par rapport à la fin de l'énoncé et par rapport à la perturbation
- Un tableau créé manuellement indiquant la fréquence des marqueurs et leurs distributions et leurs fréquences relatives aux énoncés

1. Choix des comportements transversaux

Nous avons identité 26 comportements à l'issue de l'annotation :

- 9 marqueurs audios

Étiquette	Description
Répétition partielle	Formulation d'une question comportant la partie de l'énoncé non hachurée
Question	Formulation d'une question ne comprenant aucune partie de l'énoncé
Supposition	Formulation d'une question incluant la partie de l'énoncé

	hachurée
Répétition	Répétition d'une partie non hachurée de l'énoncé (sans interaction)
Verbalisation de l'incompréhension	Formulation d'une affirmation visant à avertir l'expérimentateur de la hachure ou de l'incompréhension
Jurons	Formulation de jurons (sans interaction)
Soupir	Expiration prolongée
Rire Sourire	accompagné d'une vocalisation cyclique
Allongement syllabique	Allongement d'une ou plusieurs syllabes lors de la formulation de la rétroaction

Tableau 1: Liste et description des marqueurs audios

- 17 marqueurs visuels

Étiquette	Description
Hochement vertical	Rotation récursive de la tête de gauche à droite
Sourire	Étirement des coins des lèvres vers le haut
Avancement des lèvres	Contraction et mouvement des lèvres vers l'avant
Écarquillement des yeux	Ouverture importante des paupières
Haussement d'un sourcil	Étirement du coin interne de l'un des deux sourcils vers le haut
Plissement des yeux	Contraction des paupières
Divers lèvres 26	Macro catégorie regroupant une série de mouvements de lèvres isolés présents uniquement chez le sujet n°4
Divers tête 7	Rotation d'un seul côté de manière non récursive
Divers tête 8	Macro catégorie regroupant les mouvements de tête isolés
Contraction du menton	Contraction du menton
Étirement d'un côté des lèvres	Étirement d'un côté des lèvres
Étirement des lèvres	Étirement des deux coins des lèvres
Froncement de sourcils	Abaissement et rapprochement des coins internes des sourcils
Avancement de la tête	Mouvement de la tête vers l'avant
Divers lèvre 18	Macro catégorie regroupant des mouvements de lèvres isolés indépendamment des sujets.
Haussement des sourcils	Étirement des coins internes et externes des sourcils vers le haut
Sourire inversé	Étirement des coins des lèvres vers le bas

Tableau 2: Liste et description des marqueurs visuels

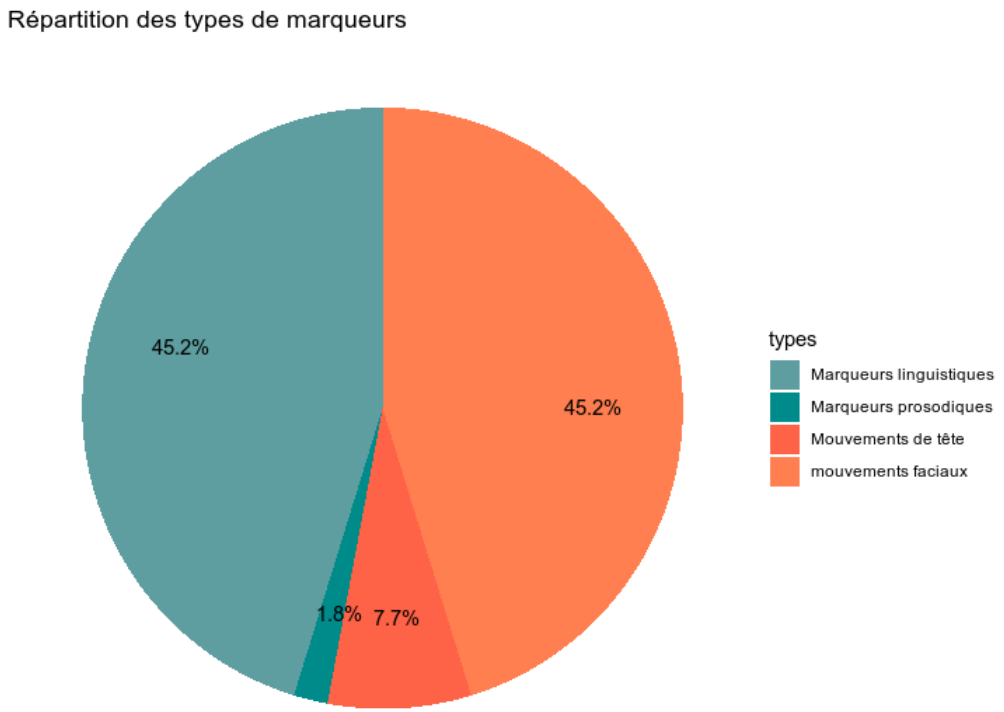


Figure 4: Répartition des types de marqueurs

Nous observerons par la suite les comportements (cf. Tableau 1 et Tableau 2) présents sur 14 sujets, sur les deux phases de communication avec et sans interaction, en excluant les phases tests. Nous les avons réparties d'abord en deux catégories selon le canal d'émission (audio ou visuel) puis en quatre grandes catégories :

- mouvements faciaux
- mouvements de tête
- marqueurs linguistiques
- marqueurs prosodiques

En observant la répartition de ces marqueurs en fonction du canal, nous remarquons une distribution plutôt homogène (cf. Figure 4).

L'importance des marqueurs visuels peut s'expliquer par le fait que la phase sans interaction n'implique pas de marqueurs linguistiques, à l'inverse de la phase avec interaction En observant de plus près la répartition des marqueurs à l'intérieur des deux modalités audio-visuelles, nous remarquons une répartition identique des marqueurs faciaux et linguistiques.

Nous nous sommes ensuite intéressés précisément à la fréquence de chacun des marqueurs. Lors de cette première analyse nous n'avons pas pris en compte les macro catégories afin de ne traiter que des comportements clairement identifiés.

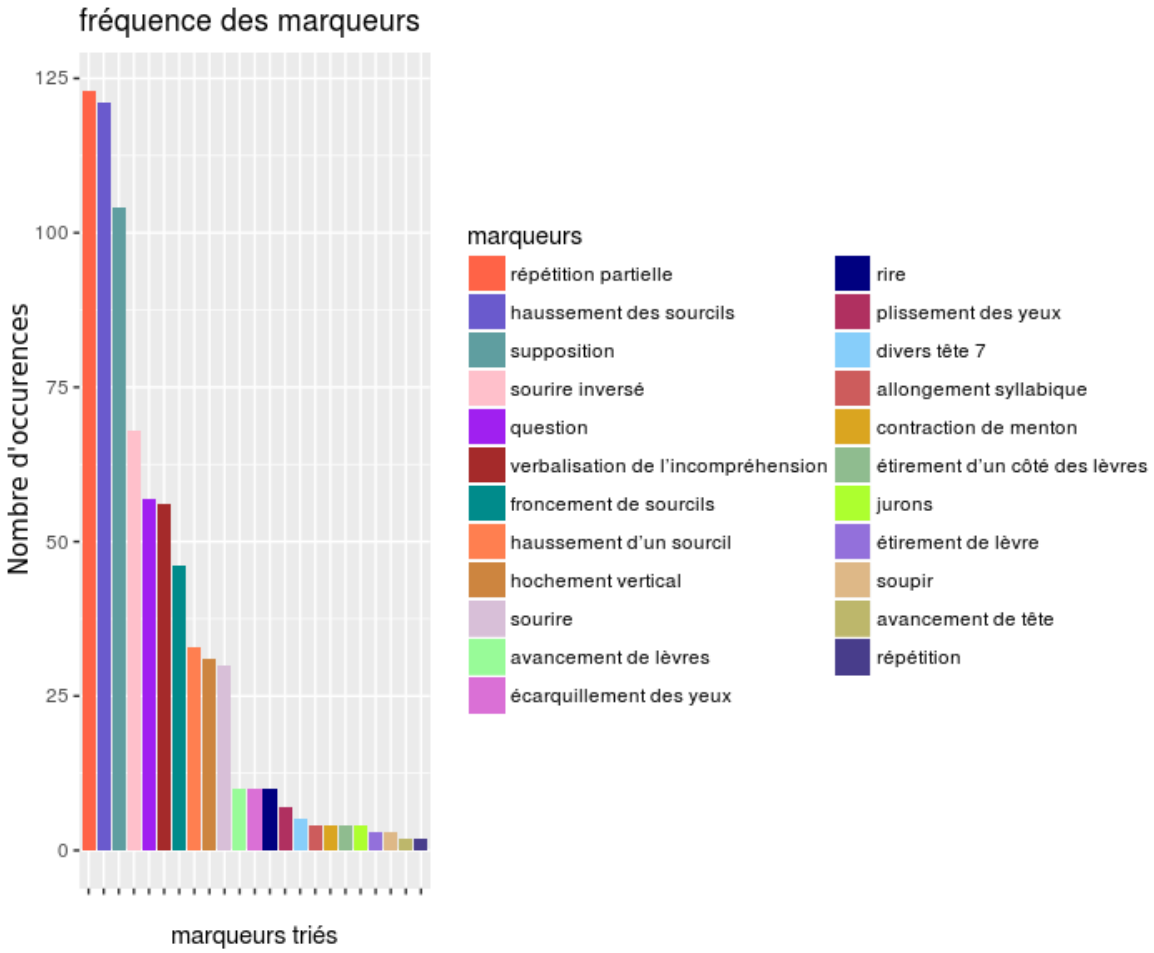


Figure 5: fréquence des marqueurs

En observant tout d'abord ces fréquences (Figure 5), nous pouvons noter une démarcation nette entre le marqueur sourire qui dépasse les 25 occurrences et le marqueur avancement de lèvres qui n'est qu'à 10 occurrences. Nous n'avons pas pris en compte les 3 macros catégories dans cette analyse. Mais la seule information sur la fréquence ne peut pas être le seul critère pour déterminer la transversalité de nos marqueurs. Notre deuxième critère sera donc la distribution des marqueurs selon les sujets (cf. Figure 6). Nous considérons qu'un comportement transversal devra être présent chez au moins la moitié de nos sujets. Les marqueurs retenus sont donc:

Marqueurs	Présence
Haussement de sourcils	13 sujets
Répétition partielle	12 sujets
Sourire inversé	12 sujets
Question	11 sujets
Supposition	11 sujets

Sourire	10 sujets
Froncement des sourcils	9 sujets
Hochement horizontal	8 sujets
Verbalisation de l'incompréhension	8 sujets

Tableau 3: marqueurs retenus

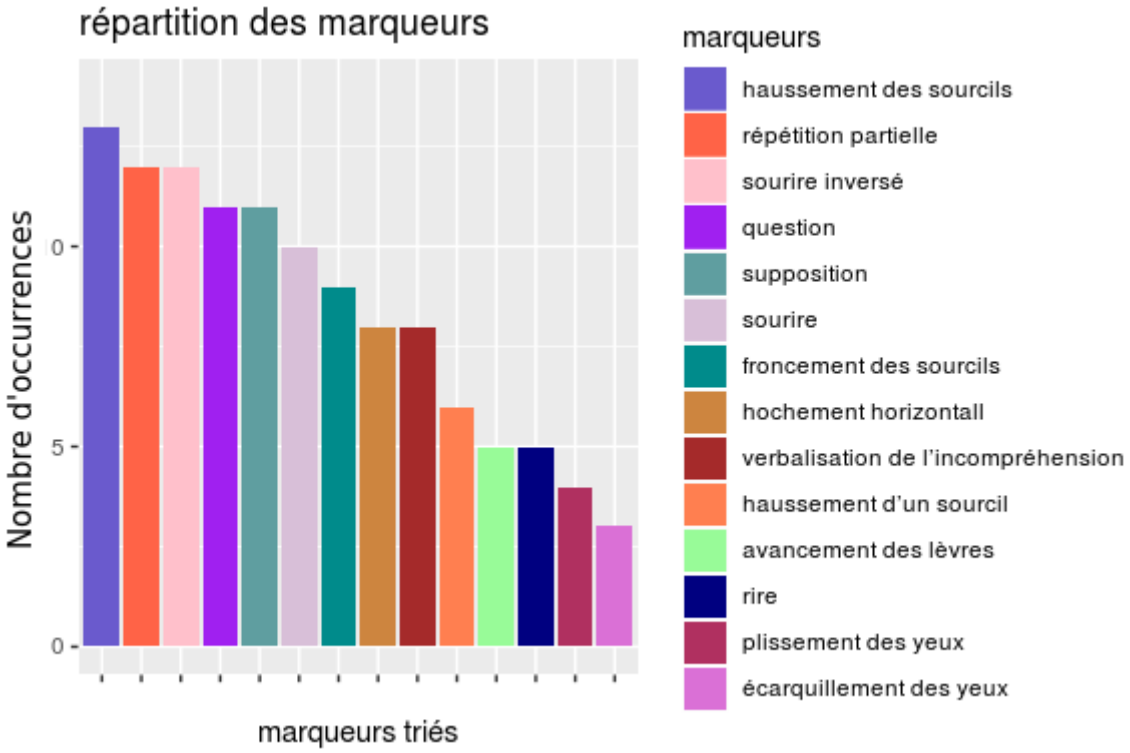


Figure 6: Répartition des marqueurs

2. Caractérisation des marqueurs retenus

Nous nous sommes ensuite penchés sur les informations de durées, de latences et de cooccurrences.

Calculs sur les durées en seconde		
Marqueurs	Moyennes	Écarts-types
Haussement des sourcils	0,719	0,692
Répétition partielle	0,897	0,380
Sourire inversé	0,876	0,759
Question	0,682	0,295
Supposition	0,903	0,532
Sourire	**2,831**	**1,833**

Froncement des sourcils	**1,937**	**1,613**
Hochement vertical	0,593	0,312
Verbalisation de l'incompréhension	1,137	0,564

Tableau 4: moyenne des durées des marqueurs

En observant d'abord les durées des marqueurs, nous pouvons voir que ceux-ci suivent la même tendance, relativement homogène, à savoir une durée moyenne oscillant entre une demi seconde et une seconde. Deux marqueurs cependant se distinguent : le sourire et le froncement de sourcils qui ont une durée moyenne largement supérieur aux autres marqueurs mais aussi une répartition très hétérogène comme en témoignent l'écart-type significativement plus haut par rapport aux autres marqueurs

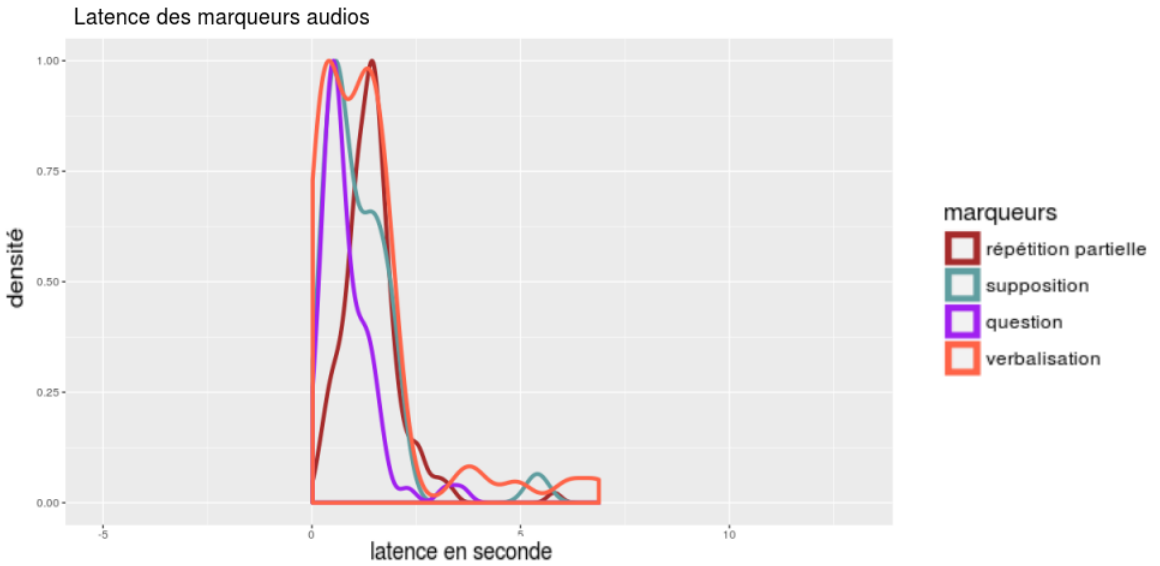


Figure 7: distribution des délais d'apparition des marqueurs audios par rapport à la fin de l'énoncé

Nous avons par la suite calculé la latence des marqueurs par rapport à la fin de l'énoncé. Une latence négative signifierait donc que le marqueur apparaît avant la fin de l'énoncé (cf. Figure 3).

Du côté des marqueurs audios, les marqueurs apparaissent en majorité entre 0 et 3 secondes après la fin de l'énoncé. Nous voyons tout de même une tendance du marqueur "verbalisation de l'incompréhension" qui s'étend après 5s et aucun marqueur produit avant 0)

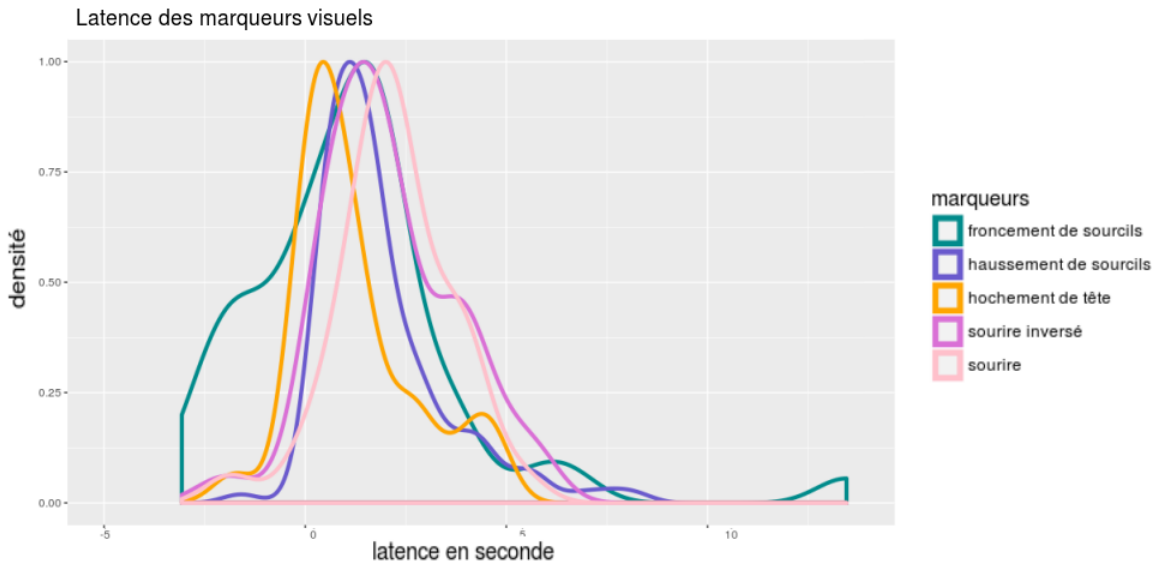


Figure 8: distribution des délais d'apparition des marqueurs visuels par rapport à la fin de l'énoncé

Du côté des marqueurs visuels (cf. Figure 8) nous remarquons en revanche que la distribution est beaucoup plus étendue et montre des valeurs de latences inférieur à 0.

Nous avons par la suite fait une calcul de cooccurrence des marqueurs retenus. Nous avons fait un calcul en uni-gramme, c'est à dire la fréquence d'apparition de deux marqueurs présent dans le même intervalle de temps, puis en bi-gramme, c'est-à-dire la fréquence d'apparition de trois marqueurs présents dans le même intervalle de temps. Cependant, ce dernier ne nous indiquait que des cas isolés.

Cooccurrence des marqueurs				
Marqueur 1	Marqueur 2	Fréquence	Pourcentage1	Pourcentage2
Hochement	Verbalisation	30	**96,77%**	**53,57%**
Haussement	Question	30	24,79%	**52,63%**
Sourire inversé	Haussement	27	**39,71%**	22,31%
Haussement	Répétition partielle	21	17,07%	17,07%
Verbalisation	Sourire	12	21,43%	**40,00%**
Verbalisation	Haussement	10	17,86%	8,26%
Sourire	Répétition partielle	10	**33,33%**	8,13%
Supposition	Froncement	9	8,65%	19,57%
Répétition partielle	Froncement	7	5,69%	15,22%
Verbalisation	Répétition partielle	7	12,50%	5,69%
Haussement	Supposition	5	4,13%	4,81%
Sourire	Haussement	5	16,67%	4,13%

Haussement	Froncement	4	3,31%	8,70%
Sourire	Question	4	13,33%	7,02%
Supposition	Sourire	4	3,85%	13,33%
Haussement	Haussement	4	3,31%	3,31%
Sourire inversé	Hochement	3	4,41%	9,68%
Froncement	Froncement	3	6,52%	6,52%
Sourire inversé	Froncement	3	4,41%	6,52%
Hochement	Répétition partielle	3	9,68%	2,44%
Haussement	Hochement	3	2,48%	9,68%
Verbalisation	Question	2	3,57%	3,51%
Verbalisation	Sourire inversé	2	3,57%	2,94%

Tableau 5: Liste des cooccurrences

En observant les cooccurrences nous notons que certains marqueurs apparaissent presque systématiquement en paire avec un autre marqueur. C'est le cas du hochement horizontal qui apparaît dans 96,77% des cas accompagnés d'une verbalisation de l'incompréhension. Le marqueur "sourire" apparaît presque systématiquement accompagné d'un marqueur oral ("question", "verbalisation de l'incompréhension" ou "répétition partielle"). Ce même phénomène est présent pour le marqueur "haussement de sourcils". Nous notons aussi une tendance du marqueur "Sourire inversé" à apparaître en cooccurrence avec un haussement de sourcils.

3. Caractérisation selon les conditions

Nous avons par la suite fait une analyse de variance sur la fréquence, la durée et le délai d'apparition des marqueurs afin de vérifier l'effet de la phase d'interaction, de la position de la perturbation et de leur interaction. Nous avons réaliser un test ANOVA à partir d'un modèle mixte (librairie lme sous R) pour la variance des délais et des durée et à partir d'une régression beta (librairie GlmmADMB sous R) pour les fréquences (les données de fréquences étaient discrètes et bornées et donc non adaptées à un modèle mixte standard).

3.1. Modulations des fréquences

Lors de la série d'analyse sur les fréquences, afin d'estimé la significativité d'une condition, nous l'avons comparée avec la condition AISP, pendant laquelle aucun marqueur n'est sensé être produit étant donné qu'aucun élément n'est présent pour perturber la compréhension.

Le marqueur "sourire" est produit avec une fréquence variant significativement en fonction de la position de la hachure dans l'énoncé (facteur Position), du degré

d'interaction avec l'expérimentateur (facteur Interaction) et de leur interaction (Phase*Position : df=3, LRatio=12.29, p=0.006 **)(cf. Figure 9). Ce marqueur semble être plutôt un marqueur d'incompréhension pour autrui puisque sa fréquence de production n'augmente pas significativement en condition non interactive (par rapport à la condition de référence AISP) mais seulement en condition interactive, en position P2 lorsque l'incompréhension est vraiment importante (+1.4 (sur 8 énoncés), p=0.002).

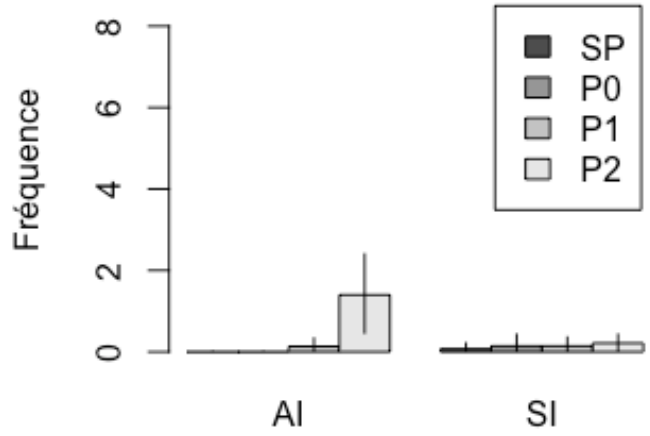


Figure 9: fréquence du marqueur sourire

Le marqueur "hochement de la tête" ne varie pas significativement en fonction du facteur Position (df=3, LRatio= 3.95, p=0.27 NS) ou du facteur Interaction (df=1, LRatio=0.30, p=0.58 NS)(cf. Figure 10). Ce marqueur ne peut donc pas, à cette étape d'analyse, être considéré comme un marqueur d'incompréhension transversal puisqu'il n'est pas davantage produit lors des conditions où il y a une hachure.

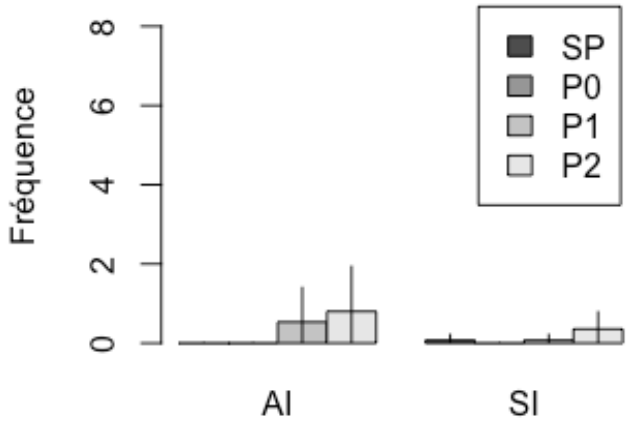


Figure 10: fréquence du marqueur hochement horizontal

Le marqueur "froncement de sourcils" est produit avec une fréquence variant significativement en fonction du facteur Position (df=3, LRatio= 19.09, p=0.0002 ***) mais non en fonction du facteur Interaction (df=1, LRatio= 0.25, p=0.61 NS). La fréquence

de ce marqueur d'incompréhension augmente significativement, par rapport à la condition de référence AISP, à la fois en situation interactive et non interactive, et peut donc être interprété comme un marqueur d'incompréhension pour soi. Cependant, l'augmentation de sa fréquence de production n'est significative que dans les positions P1 et P2 de la hachure (respectivement +0.5 (sur 8 énoncés), p=0.037 et +0.6, p=0.022*).

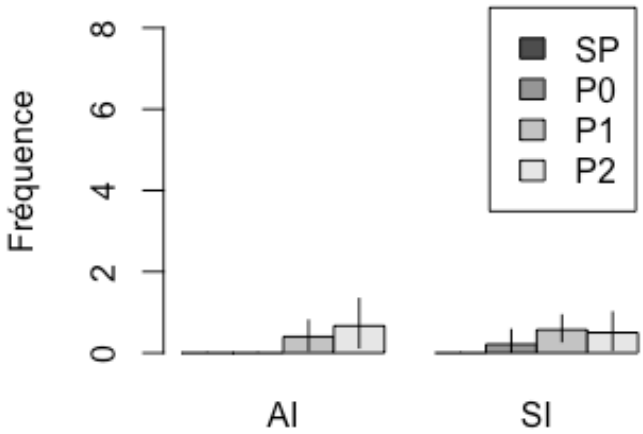


Figure 11: fréquence du marqueur froncement de sourcils

Le marqueur "haussement de sourcils" est produit avec une fréquence variant significativement en fonction du facteur Position (df=3, LRatio= 11.25, p=0.01*) mais non en fonction du facteur Interaction (df=1, LRatio= 0.28, p=0.59 NS). La fréquence de ce marqueur d'incompréhension augmente significativement, par rapport à la condition de référence, à la fois en situation interactive et non interactive, et peut donc être interprété comme un marqueur d'incompréhension pour soi. Cependant, l'augmentation de sa fréquence de production n'est significative que dans les positions P1 et P2 de la hachure (+1,6 p=0,039*).

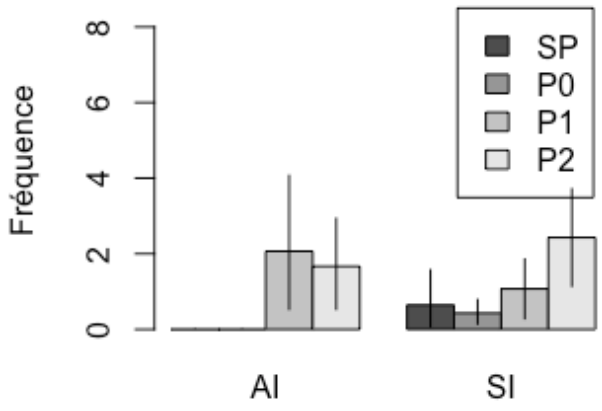


Figure 12: fréquence du marqueur haussement de sourcils

Le marqueur "sourire inversé" est produit avec une fréquence variant significativement en fonction de la position de la hachure dans l'énoncé, du degré d'interaction avec l'expérimentateur et de leur interaction (df=3, LRatio=8,56, p=0.036*) (cf. Figure 13). Ce marqueur semble être plutôt un marqueur d'incompréhension pour soi puisque sa fréquence de production augmente significativement en condition non interactive par rapport au niveau de référence. De plus il est produit avec un fréquence significative en position P2 lorsque l'incompréhension est vraiment importante (+2.0, p=0.002**).

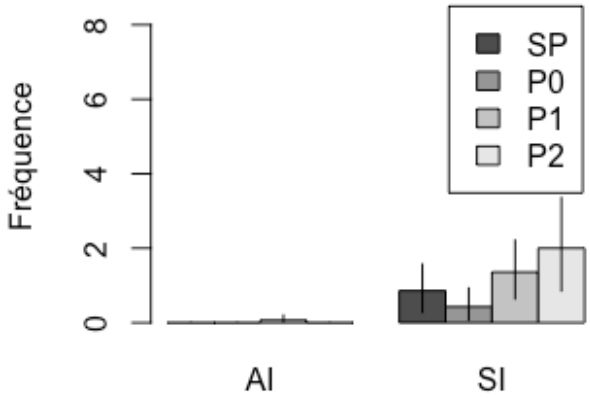


Figure 13: fréquence du marqueur sourire inversé

Le marqueur "répétition partielle" est produit avec une fréquence variant significativement en fonction de la position de la hachure dans l'énoncé, du degré d'interaction avec l'expérimentateur et de leur interaction (df=3, LRatio=41.51, p=0.0001 ***)(cf. Figure 14). Ce marqueur semble être plutôt un marqueur d'incompréhension pour autrui puisque sa fréquence de production augmente significativement en condition interactive. De plus il est produit avec un fréquence significative en position P2 lorsque l'incompréhension est vraiment importante (+7.1, p=0.000***).

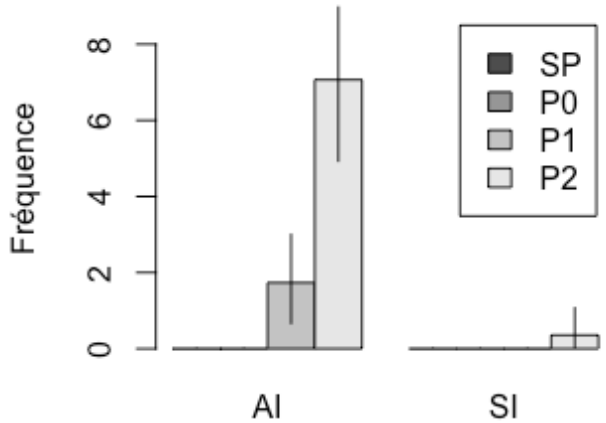


Figure 14: fréquence du marqueur répétition partielle

Le marqueur "verbalisation de l'incompréhension" ne varie pas significativement en fonction du facteur Position (df=1, LRatio= 1.68, p=0.19 NS) ou du facteur Interaction (df=3, LRatio=2.61, p=0.46 NS)(cf. Figure 15). À ce stade d'analyse, ce marqueur ne peut donc pas être considéré comme un marqueur d'incompréhension transversal puisqu'il n'est pas produit de manière significative lors des conditions où il y a une hachure.

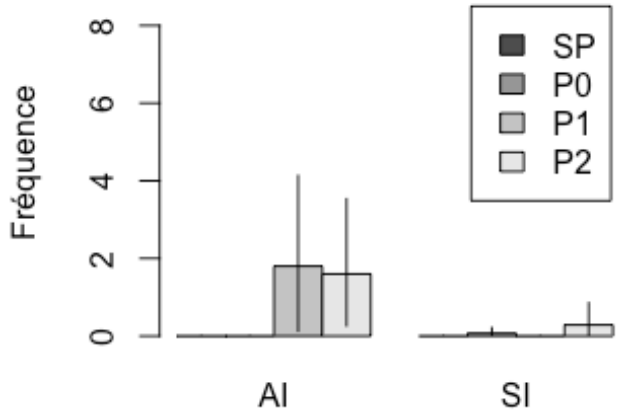


Figure 15: fréquence du marqueur verbalisation de l'incompréhension

Le marqueur "supposition" est produit avec une fréquence variant significativement en fonction de la position de la hachure dans l'énoncé, du degré d'interaction avec l'expérimentateur et de leur interaction (df=3, LRatio=19.88, p=0.0001 ***)(cf. Figure 16). Ce marqueur semble être plutôt un marqueur d'incompréhension pour autrui puisque sa fréquence de production augmente significativement en condition interactive (par rapport à la condition de référence AISP). De plus il est produit avec une fréquence significative en position P1 et P2 (respectivement +4.1, p=0.000*** et +2.8s, p=0.007***).

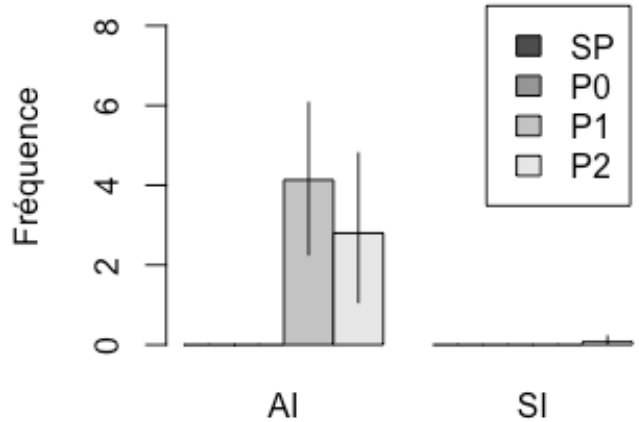


Figure 16: fréquence du marqueur supposition

Enfin le marqueur "question" est produit avec une fréquence variant significativement en fonction de la position de la hachure dans l'énoncé, du degré d'interaction avec l'expérimentateur et de leur interaction (df=3, LRatio=26,37, p=0.0001 ***)(cf. Figure 17). Ce marqueur semble être plutôt un marqueur d'incompréhension pour autrui puisque sa fréquence de production augmente significativement en condition interactive. De plus il est produit avec un fréquence significative en position P1 lorsque l'incompréhension est modérée (+3.7, p=0.000***).

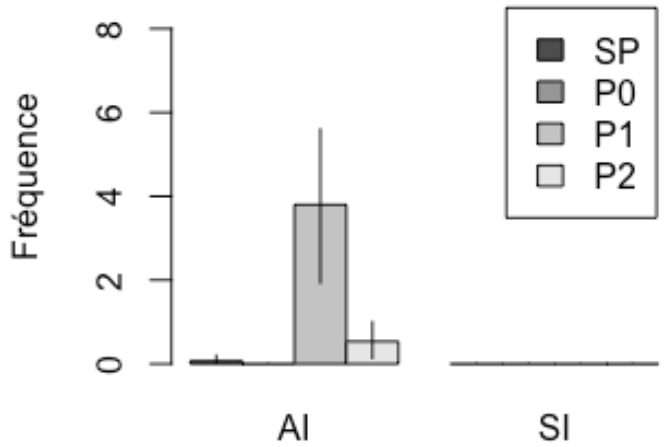


Figure 17: fréquence du marqueur question

À la fin de cette analyse, les marqueurs "hochement horizontal" et "verbalisation de l'incompréhension" peuvent être laissé de côté car leurs fréquences en situation perturbée n'est pas suffisamment significative pour qu'ils soient considérés réellement comme des marqueurs d'incompréhension transversaux, ces marqueurs ne seront donc pas retenus pour le reste des analyses. Les marqueurs "froncement de sourcils", "haussement de sourcils" et "sourire inversé" peuvent être considérés comme des marqueurs pour soi du fait que leur fréquence de production est significativement plus grande en situation non interactive et en présence de hachures, fréquence particulièrement forte en position P2 uniquement pour le sourire inversé, en position P1 et P2 pour le froncement de sourcils sans cependant de différence significative entre les positions chez ce dernier et en positions P1 et P2 combinées sans l'effet de P0 pour le haussement de sourcils. Cependant, le fait que ces deux derniers marqueurs soient présents dans les deux situations d'interaction, des analyses complémentaires mériteraient d'être faites. Les marqueurs "sourires", "répétition partielle", "supposition" et "question" peuvent être considérés comme des marqueurs d'incompréhension pour autrui, du fait que leur production est significativement plus fréquente en présence de hachures et seulement en situation interactive. L'augmentation de cette fréquence de production est significative uniquement en P2 pour le sourire et la

répétition partielle, en P1 et P2 pour la supposition sans différence significative entre les positions, et seulement en P1 pour question.

3.2. Modulations des durées

Pour cette partie, étant donné que certains marqueurs ne sont jamais produits dans certaines conditions, nous avons recodé un facteur condition à huit niveaux au lieu des deux facteurs croisés phase et position en ne prenant en compte que les conditions dans lesquelles le marqueur est produit. Ces huit niveaux correspondent à :

- 1 - Avec Interaction Sans Perturbation (AISP)
- 2 - Avec Interaction Perturbé en position 0 (AIP0)
- 3 - Avec Interaction Perturbé en position 1 (AIP1)
- 4 - Avec Interaction Perturbé en position 2 (AIP2)
- 5 - Sans Interaction Sans Perturbation (SISP)
- 6 - Sans Interaction Perturbé en position 0 (SIP0)
- 7 - Sans Interaction Perturbé en position 1 (SIP1)
- 8 - Sans Interaction Perturbé en position 2 (SIP2)

Nous avons ensuite réalisé des tests post-hoc pour examiner le contraste plus spécifique entre certaines conditions en appliquant des corrections de Bonferroni pour les comparaisons multiples.

La durée des marqueurs "sourire", "froncement de sourcils", "haussement de sourcils" et "sourire inversé" n'est pas significativement modulée en fonction des conditions (respectivement df=5, LRatio=5.94, p=0.31 NS ; df=7, LRatio=8.17, p=0.32 NS ; df=5, LRatio=7.18, p=0.21 NS ; df=4, LRatio=4.19, p=0.38 NS). Ces marqueurs d'incompréhension sont donc soit "binaire" (présents vs. absents) ou bien modulés avec le degré d'incompréhension sur d'autres dimensions (délai de production, amplitude du mouvement).

La durée de M12 (répétition partielle), M16 (assomption) et M22 (question) est significativement modulée en fonction des conditions (respectivement df=2, LRatio=16.93, p=2e-04 *** ; df=2, LRatio= 10.13, p=0.006 ** ; df=2, LRatio= 11.21, p=0.004 **), sans pour autant qu'une tendance générale puisse être dégagée entre ces trois marqueurs :

- En condition AI, les durées de M12 et M22 sont plus courtes pour des hachures en position P2 qu'en position P1 (respectivement -187 +/- 71ms, p=0.017, * ; -882 +/- 27 ms, p=0.002 **), tandis que la durée de M16 augmente (significativement dans tous les cas : +307 +/- 94ms, p=0.002 **).
- Le fait d'interagir avec l'expérimentateur (AI vs. SI) n'affecte pas significativement la durée des marqueurs M16 et M22 (respectivement -410 +/- 483 ms, p=0.63 NS ; (-176 +/- 98 ms, p=0.14 NS). En revanche, le marqueur M12 est significativement plus court en condition interactive (527 +/- 156 ms sur la durée moyenne du marqueur en position P2, p=0.001 **).

La durée de ces marqueurs semble être un indicateur du degré d'incompréhension

3.3. Modulations des délais d'apparition

Dans cette partie d'analyse, nous avons toujours recours à des modèles mixtes en substituant les effets de la position et de la phase par un facteur condition à huit niveaux. Nous analysons le délai d'apparition des marqueurs par rapport à deux repères :

- le délai d'apparition par rapport à la fin de l'énoncé
- le délai d'apparition par rapport au début de la hachure

Tout d'abord les deux marqueurs "sourire" et "sourire inversé" montrent un délai de production qui ne varie pas significativement en fonction de la condition (respectivement df=5, LRatio=10.16, p=0.07 NS ; p>0.9 pour la variation de ce délai au sein de la condition interactive pour le marqueur "sourire inversé"), de 2.271 +/- 0.320 s en moyenne par rapport à la fin de l'énoncé pour le marqueur "sourire" et de 2.131 +/- 0.319 s en moyenne pour le "sourire inversé").

Le marqueur "froncement de sourcils" varie significativement en fonction de la condition (df=7, LRatio= 23.21, p=0.002 **).

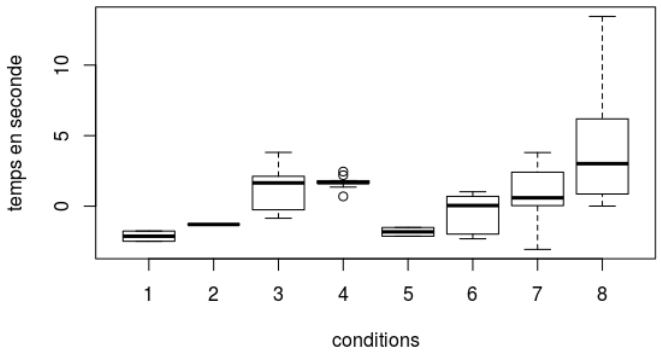


Figure 18: délai d'apparition du marqueur froncement de sourcils par rapport à la fin de l'énoncé

Il n'y a pas de différence significative entre les phases avec et sans interaction (-1.1 +/- 0.9s, p=0.60 NS) mais nous observons une tendance à produire ce marqueur de façon croissante en fonction du degré d'incompréhension. C'est à dire, plus l'information hachurée sera importante plus le délai d'apparition de ce marqueur sera grand. Nous obtenons donc un délai significatif pour la condition P2 qui apparaît en moyenne 2.0s +/- 0.7s après la fin de l'énoncé (p=0.023*)(cf. Figure 18). On remarque également que ce marqueur a tendance à être produit avant la fin de l'énoncé, ce qui nous pousse à poursuivre les analyses sur le délai par rapport au début de la hachure.

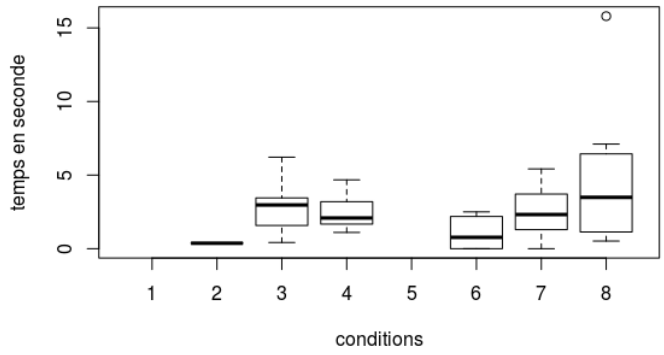


Figure 19: délai d'apparition du marqueur froncement de sourcils par rapport au début de la hachure

Le graphique ci-dessus montre une tendance croissante de ce marqueur par rapport au degré de perturbation mais celle-ci n'est pas significative. En d'autres termes, le délai de production de ce marqueur ne varie pas significativement de 2.3s +/-0.5s en moyenne (p=1 NS).

Le délai d'apparition du marqueur "haussement de sourcils" ne varie pas significativement en fonction de la phase (240 +/- 322ms, p=0.83 NS)(cf. Figure 20). En revanche, on remarque un délai d'apparition plus grand en position P2 (1.2ms) qu'en position P1 (ET=283ms, p<1e-04 ***).

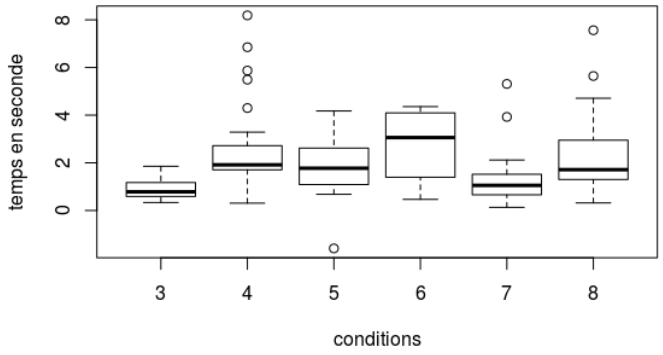


Figure 20: délai d'apparition du marqueur haussement de sourcils par rapport à la fin de l'énoncé

La différence entre P1 et P2 nous amène à pousser les analyses plus loin. La différence entre P1 et P2 n'est plus significative (303 +/- 278ms, p=0,64)(cf Figure 21), on suppose donc que l'apparition de ce marqueur est constante par rapport à la hachure et ne dépend pas de la fin de l'énoncé.

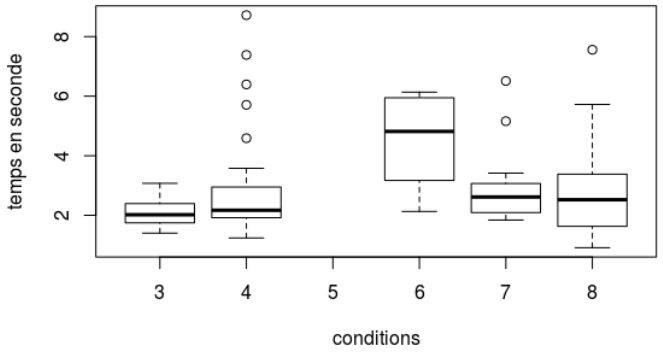


Figure 21: délai d'apparition du marqueur haussement de sourcils par rapport au début de la hachure

Le délai d'apparition du marqueur "répétition partielle" varie significativement en fonction de la position (cf. Figure 22). Ce marqueur a tendance à être produit plus tard en P2 (490ms) qu'en P1(ET=150ms, p=0.002**).

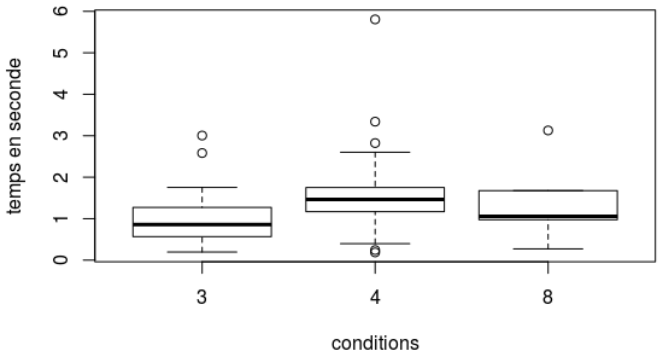


Figure 22: délai d'apparition du marqueur répétition partielle par rapport à la fin de l'énoncé

En poussant les analyses sur le délai d'apparition par rapport à la hachure, contrairement à ce qu'on aurait pu attendre, ce marqueur est produit avec un délai plus rapide suite à une perturbation en P2 qu'en P1 (464 +/-170ms, p=0.013*)(cf. Figure 23). Nous nous confrontons à un biais de notre expérience. Dans les cas où la hachure se trouve en P2, celle-ci se confond avec la fin de l'énoncé. En d'autres termes, dans cette condition la hachure tronquera les énoncés. Nous nous retrouvons donc avec des cas où la fin de l'énoncé arrive avant la hachure. Malgré cela, le fait que ce marqueur arrive plus rapidement en P2 qu'en P1 laisse supposer que cet élément soit indicateur du degré d'incompréhension.

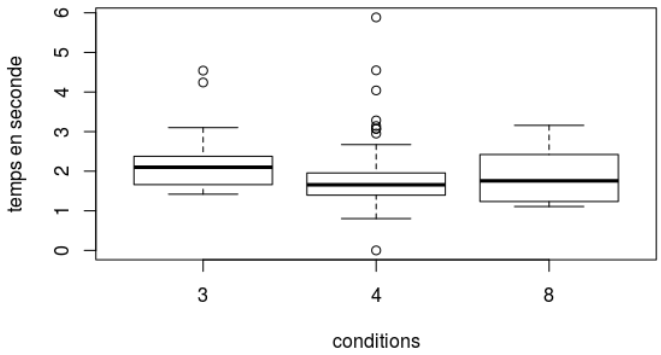


Figure 23: délai d'apparition du marqueur répétition partielle par rapport au début de la hachure

Le marqueur "supposition" apparaît en moyenne plus tard après une hachure en position P2 qu'en position P1 en condition avec interaction (910ms, ET=136ms, p<1e-10***).

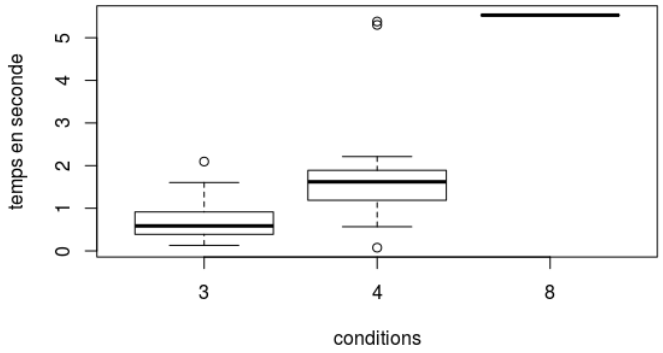


Figure 24: délai d'apparition du marqueur supposition par rapport à la fin de l'énoncé

En poussant les analyses sur le délai par rapport à la hachure, la différence entre P1 et P2 n'est plus significative avec une émission d'1.9 +/-0.1s en moyenne après la hachure(p=0.57 NS).

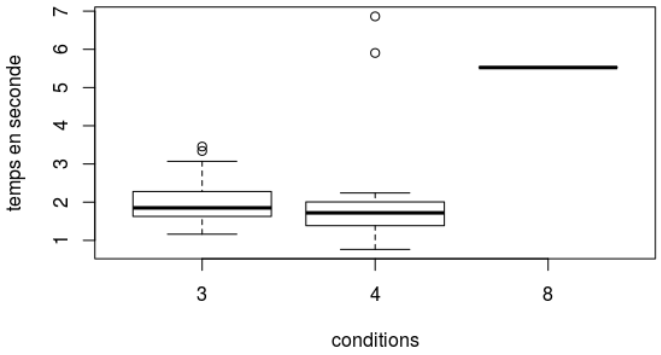


Figure 25: délai d'apparition du marqueur supposition par rapport au début de la hachure

Comme les autres marqueurs, le marqueur "question" est produit significativement plus tard en P2 (494ms) qu'en P1 en condition interactive (p=0.018*).

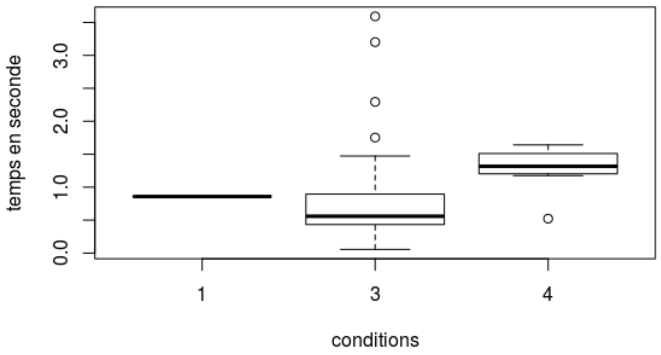


Figure 26: délai d'apparition du marqueur question par rapport à la fin de l'énoncé

En poussant les analyses sur le délai par rapport à la hachure, comme pour les autres marqueurs, nous perdons cette significativité entre P1 et P2 avec une apparition moyenne de 288 +/-262ms (p=0.27 NS).

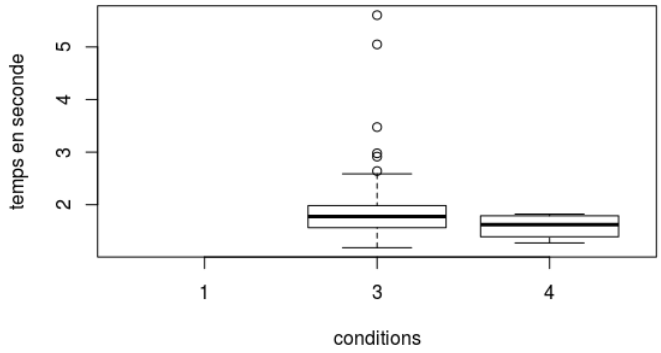


Figure 27: délai d'apparition du marqueur question par rapport au début de la hachure

Finalement pour la majorité de ces marqueurs "froncement de sourcil", "haussement de sourcils", "supposition" et "question" un délai plus long, par rapport à la fin de l'énoncé, est observé suite à une incompréhension en position P2 qu'en P1. Cette différence correspond en fait à un délai relativement constant entre le début de la hachure et la production du marqueur . En d'autres termes, ces marqueurs d'incompréhension ne sont pas "produits" exactement au moment même de la hachure, mais avec certain délai, assez constant, par rapport à cette hachure ("supposition" : 1.9 +/- 0.1 s ; "question" : 1.8 +/- 0.2 s). Le fait qu'ils soient réalisés et perceptibles après la fin de l'énoncé, la plupart du temps, ne signifie pas que le sujet attende la fin de l'énoncé pour les générer. Ces marqueurs semblent être émis en réaction à la hachure, sans attendre la fin de l'énoncé, mais avec une certaine latence.

Pour un de ces marqueurs, "répétition partielle", le délai plus long observé en position P2 qu'en P1 (490 +/- 150 ms, p=0.002 **) ne correspond pas à un temps de latence relativement fixe dans l'émission du marqueur par rapport au début de la hachure : le délai d'émission de ce marqueur par rapport à l'instant de hachure est plus court en position P2 qu'en position P1. On peut donc envisager que pour ce marqueur, le délai d'émission soit un indicateur du degré d'incompréhension.

Chapitre 6. Analyses sur les Unités d'Action Faciale

OpenFace (Baltrusaitis et al. 2018) est un outil de détection de mouvements faciaux et de tête développé en C++. Nous avons appliqué cet outil à l'intégralité de notre corpus et nous avons extrait les données relatives aux unités d'action faciales (AU) dans nos trois intervalles de temps (cf. Figure 3). Nous avons ensuite synthétisé ces données en calculant la moyenne, l'écart-type et le maximum d'activation dans ces intervalles puis nous avons ajouté quatre colonnes correspondant à nos quatre marqueurs visuels retenus dans lesquelles nous avons précisé si pour chaque intervalle ces marqueurs étaient produits ou non.

Nous avons appliqué comme précédemment un modèle mixte à ces unités d'action faciales afin de savoir lesquelles s'appliquent le mieux à nos marqueurs faciaux retenus. De façon à contourner le problème des distributions bimodales observées pour chaque FAU, avec un gros pic vers 0 ("bruit" dans les mouvements du visage) et un deuxième pic de moindre amplitude à des valeurs plus élevées (correspondant probablement aux mimiques faciales produites), nous avons choisi de comparer la valeur des unités d'action faciale lorsqu'il n'y a pas de marqueur émis, avec la valeur des unités d'action faciale lorsqu'un marqueur est produit, en fonction des conditions expérimentales (phase et position). Nous avons codé un facteur condition à 9 niveaux :

- C0 (référence) : pour tous les énoncés où le marqueur analysé n'a pas été produit, quelle que soit la condition expérimentale (phase, position)
- C1 à C8 : pour les énoncés en conditions de AISP à SIP2 où le marqueur analysé a été produit

Nous avons vu précédemment que certains marqueurs pouvaient être produits avant la fin de l'énoncé, nous avons donc fait les analyses sur les maximums d'activation dans les intervalles I1 et I2.

1. Marqueur haussement de sourcils

Pour le marqueur "haussement de sourcils", l'analyse se fait sur 7 niveaux (C0 + C3 à C8) du fait que ce marqueur n'est pas produit en AISP et AIP0. Après une série d'analyse, les unités d'action 1, 2, 6, 7, 9, 10, 12, 14, 15, 17, 23, 25, 26 et 45 montrent une augmentation significative par rapport à leur niveau de référence lorsque ce marqueur est

produit. Les unités d'action 6, 7 10, 12, 14, 25, 26 et 45 ne montrent une augmentation significative que pour certaines conditions mais pas de manière générale, ce qui laisse penser que ces unités ne sont pas reliées à notre marqueur.

L'unité d'action 1 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions ($p<0.001$**) et une augmentation moyenne de 1.02 +/- 0.1. Nous trouvons aussi une différence significative entre les phases d'interaction AI(P1 et P2) et SI(P1 et P2) ($p=0.02$*) et une augmentation de 0.5 +/- 0.2. En revanche il n'y a pas d'augmentation significative entre les différentes positions P1 et P2.

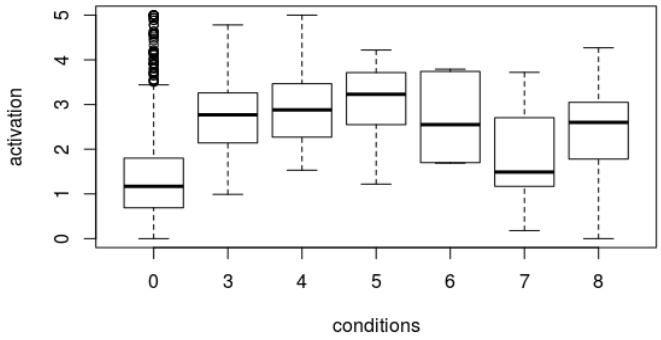


Figure 28: Maximum d'activation de l'unité d'action 1

L'unité d'action 2 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions ($p< 0.001$**) avec une augmentation moyenne de 0.7 +/- 0.097. Nous ne trouvons en revanche pas de différence significative entre les phases d'interaction AI(P1 et P2) et SI(P1 et P2) ni entre les différentes positions.

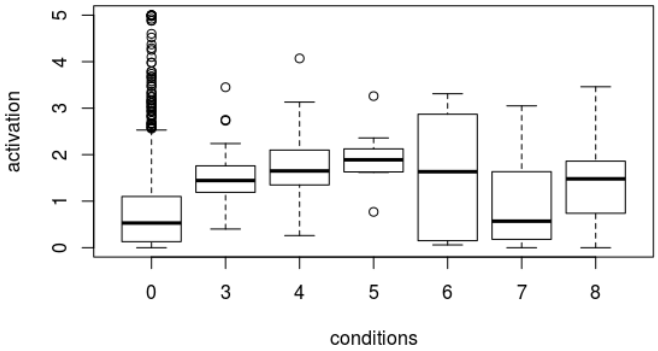


Figure 29: Maximum d'activation de l'unité d'action 2

L'unité d'action 9 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p=0.00147**) avec une augmentation moyenne de 0.22 +/- 0.06. Nous ne trouvons de différence significative ni entre les phases d'interaction AI(P1 et P2) et SI(P1 et P2) ni entre les différentes positions.

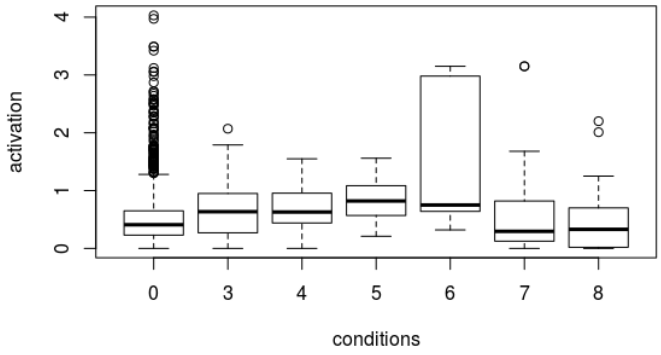


Figure 30: Maximum d'activation de l'unité d'action 9

L'unité d'action 15 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p<0.001**) avec une augmentation moyenne de 0.044 +/- 0.09. Nous ne trouvons pas de différence significative entre les phases d'interaction AI(P1 et P2) et SI(P1 et P2) ni entre les différentes positions.

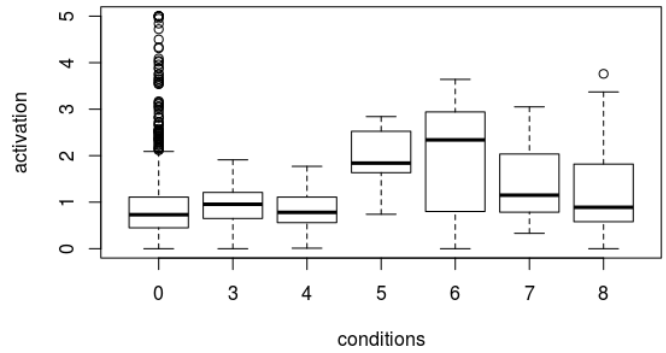


Figure 31: Maximum d'activation de l'unité d'action 15

L'unité d'action 17 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p<0.001**) avec une augmentation moyenne de 0.39 +/- 0.8. Nous ne trouvons pas une différence significative entre les phases d'interaction AI(P1 et P2) et SI(P1 et P2) ni entre les différentes positions.

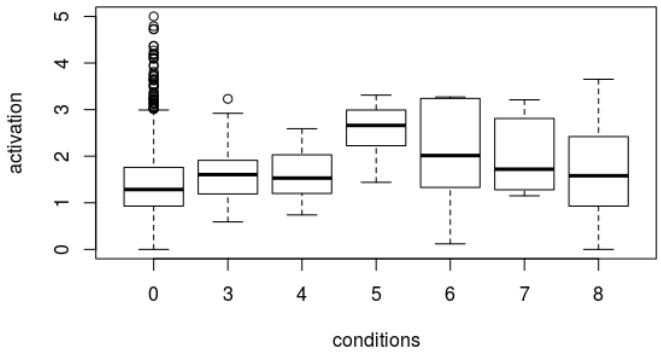


Figure 32: Maximum d'activation de l'unité d'action 17

L'unité d'action 23 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions ($p<0.001$**) avec une augmentation moyenne de 0.36 +/- 0.09. Nous ne trouvons pas une différence significative entre les phases d'interaction AI(P1 et P2) et SI(P1 et P2) ni entre les différentes positions.

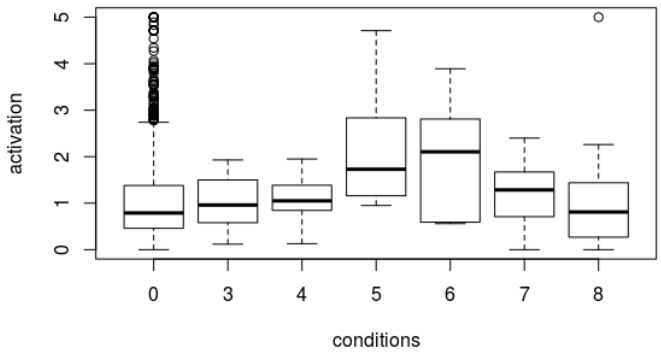


Figure 33: Maximum d'activation de l'unité d'action 23

L'unité d'action 45 ne montrait une augmentation significative que dans certaines conditions, en revanche elle montre une différence significative d'activation entre les position P0 et P1 avec une augmentation de 0.071 +/- 0.31 ($p= 0.006$**).

Les unités d'action 6, 7, 12, 14, 25 et 26 ne montraient une augmentation significative que dans certaines conditions en revanche, elles montrent toutes une différence significative entre AI (P1 et P2) et SI (P1 et P2). Deux explications sont possibles. Ce marqueur peut avoir deux rôles distincts qui se manifestent dans l'intensité de sa réalisation, lever les sourcils pour poser une question par exemple. Cette différence peut aussi bien être un artefact lié à la situation interactionnelle.

2. Marqueur froncement de sourcils

Pour le marqueur "froncement de sourcils", l'analyse se fait sur 9 niveaux : C0 + C1 à C8 du fait que ce marqueur est produit dans toutes les conditions expérimentales. Après une série d'analyse, seules les unités d'action 4, 9 et 15 montrent une augmentation significative par rapport à leur niveau de référence lorsque ce marqueur est produit. L'unité d'action 15 ne montre une augmentation significative que pour certaines conditions mais pas de manière générale, ce qui laisse penser que cette unité n'est pas reliée à notre marqueur.

L'unité d'action 4 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions ($p<0.001$**) avec une augmentation moyenne de 0.9 +/- 0.11. Nous observons aussi une augmentation significative en P1 et P2 de 0.73 +/- 0.27 ($p=0.0346$*).

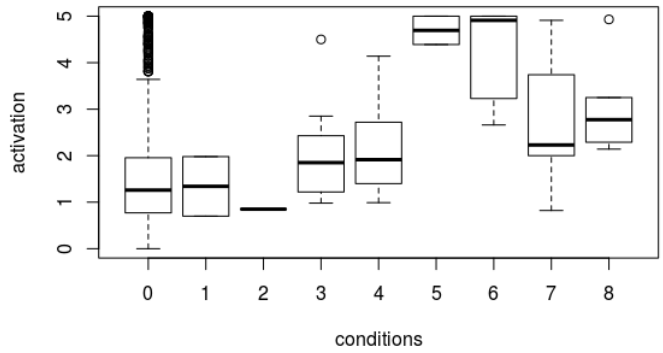


Figure 34: Maximum d'activation de l'unité d'action 4

L'unité d'action 9 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions ($p=0.038$*) avec une augmentation moyenne de 0.29 +/- 0.1. En revanche il n'y a pas d'augmentation significative entre les différentes positions P1 et P2.

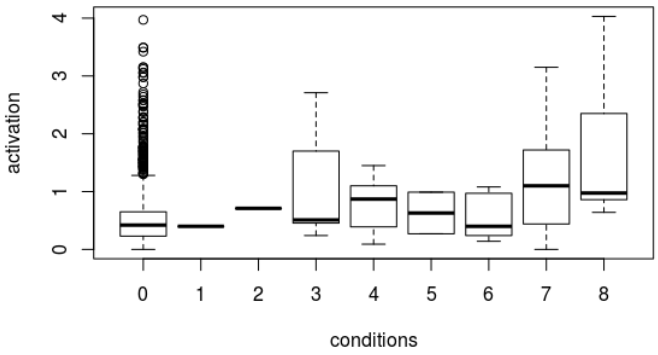


Figure 35: Maximum d'activation de l'unité d'action 9

3. Marqueur sourire inversé

Pour le marqueur "sourire inversé", l'analyse se fait sur 6 niveaux : C0 + C3 puis C5 à C8 du fait que ce marqueur n'est pas produit dans les conditions AISP, AIP0 et AIP2. Après une série d'analyses, les unités d'action 7, 15, 17, 20 et 25 montrent une augmentation significative par rapport à leur niveau de référence lorsque ce marqueur est produit. Les unités d'action 7, 20 et 25 ne montrent une augmentation significative que pour certaines conditions et ne semblent donc pas reliées à notre marqueur.

L'unité d'action 15 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p=1e-04***) avec une augmentation moyenne de 1.19 +/- 0.16. Une différence significative est également observée entre les positions SP et P0 en phases sans interaction (p=0.004**) avec une augmentation de 1.03 +/- 0.32 puis entre les positions P1 et P0 en phase sans interaction (p<0.001**) avec une augmentation de 1.3 +/- 0.3.

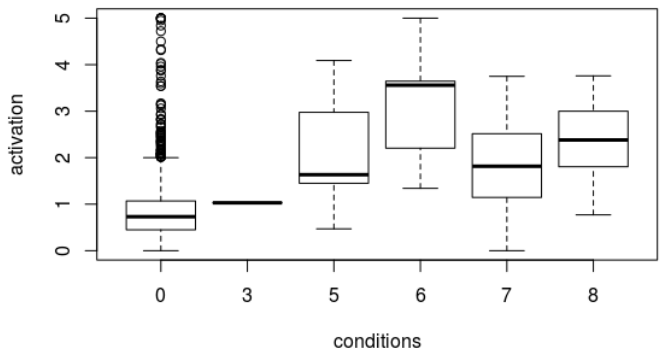


Figure 36: Maximum d'activation de l'unité d'action 15

L'unité d'action 17 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p=1e-04***) avec une augmentation moyenne de 0.8 +/- 0.15. Une différence significative est également observée entre les positions SP et P0 en phase sans interaction (p=0.013*) avec une augmentation de 0.87 +/- 0.3 puis entre les positions P1 et P0 en phase sans interaction (p=0.004**) avec une augmentation de 0.9 +/- 0.2.

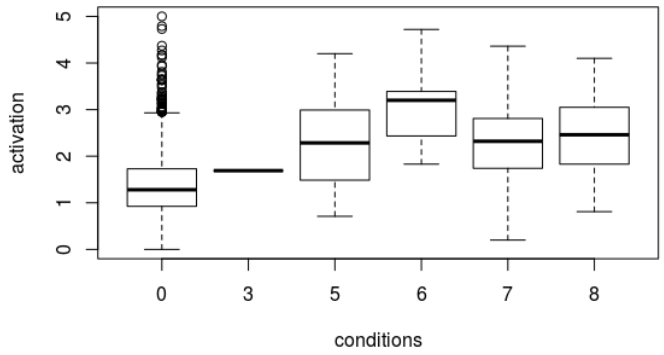


Figure 37: Maximum d'activation de l'unité d'action 17

4. Marqueur sourire

Pour le marqueur "sourire", l'analyse se fait sur 7 niveaux : C0 + C3 à C8 du fait qu'il n'existe aucune occurrence de ce marqueur en condition AISP et AIP0. Après une série d'analyses, seules les unités d'action 1, 2, 6, 7, 9, 10, 12, 14 et 25 montrent une augmentation significative par rapport à leur niveau de référence lorsque le marqueur sourire est produit. Les unités d'action 1, 2, 7 et 25 ne montrent une augmentation significative que pour certaines conditions mais pas de manière générale, ce qui laisse penser que ces trois unités ne sont pas reliées à notre marqueur.

L'unité d'action 6 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p<0.001**) et une augmentation moyenne de 1.067 +/- 0.157. En revanche il n'y a pas d'augmentation significative entre les différente position P1 et P2

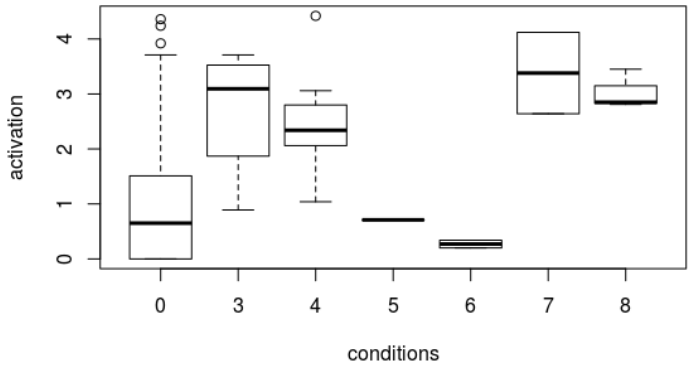


Figure 38: Maximum d'activation de l'unité d'action 6

L'unité d'action 9 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions avec (p=0.0469*) avec une augmentation moyenne de 0.3 +/- 0.13. En revanche il n'y a pas d'augmentation significative entre les différentes positions P1 et P2

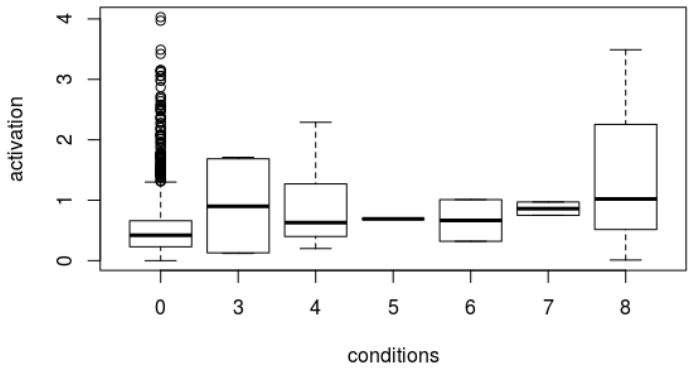


Figure 39: Maximum d'activation de l'unité d'action 9

L'unité d'action 10 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p=0.001**) et une augmentation moyenne de 0.63 +/- 0.16. En revanche il n'y a pas d'augmentation significative entre les différentes positions P1 et P2.

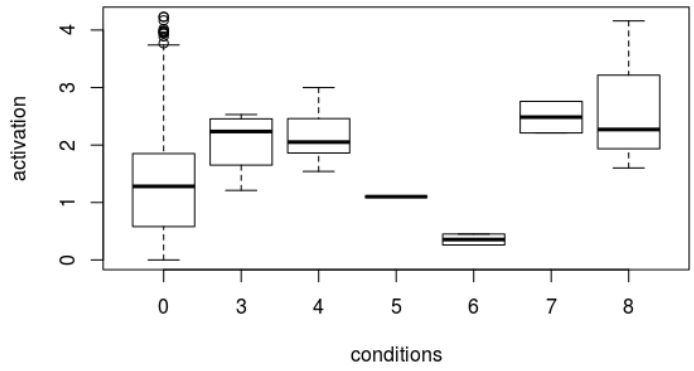


Figure 40: Maximum d'activation de l'unité d'action 10

L'unité d'action 12 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p=0.001**) avec une augmentation moyenne de 1.48 +/- 0.18. En revanche il n'y a pas d'augmentation significative entre les différentes positions P1 et P2.

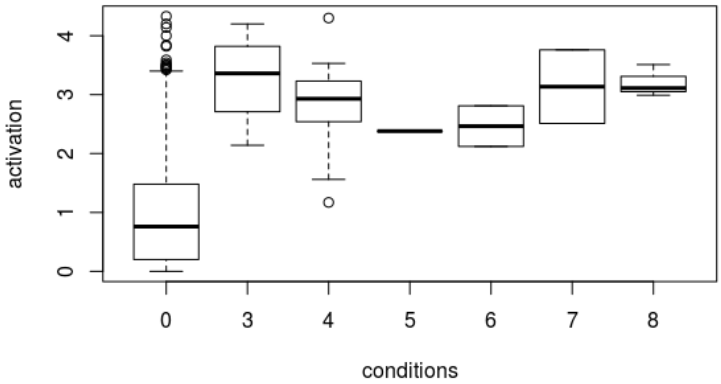


Figure 41: Maximum d'activation de l'unité d'action 12

L'unité d'action 14 augmente significativement par rapport à son niveau de référence de manière générale pour toutes les conditions (p<0.001**) avec une augmentation moyenne de 0.93 +/- 0.18. En revanche il n'y a pas d'augmentation significative entre les différentes positions P1 et P2.

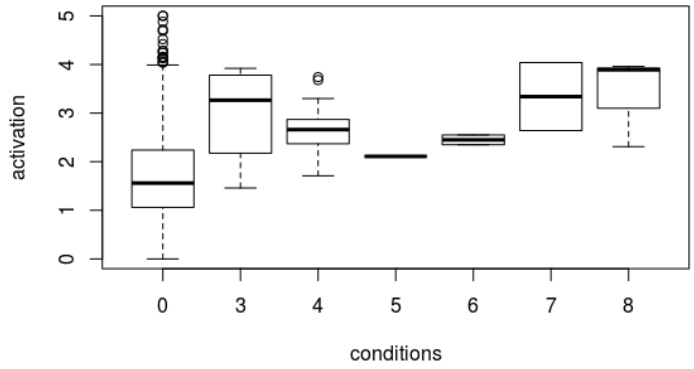


Figure 42: Maximum d'activation de l'unité d'action 14

Conclusion

Nul ne conteste aujourd'hui l'importance du non verbal dans les études sur la communication. Des auteurs tels que Ekman, Bavelas ou Allwood ont largement analysé les apports gestuels et il est reconnu aujourd'hui, par soucis d'écologie, que ce sont des éléments à prendre en compte dans les études impliquant l'interaction. Les développements de plus en plus importants dans le domaine du dialogue homme-machine par exemple, impliquent nécessairement une attention sur celui avec qui on interagit. Comment va t-il réagir et comment se réadapter à ces réactions? En observant la littérature dans ce domaine, il est frappant de constater que celui qui écoute parler, et non pas celui qui parle, n'est presque jamais pris en considération.

Nous essayons donc de savoir dans ce travail de quelles manières le récepteur d'un message exprime son incompréhension. Nous avons de plus orienté notre travail vers deux aspects :

- De quelle manière la situation interactionnelle influence t-elle la réaction d'un sujet face à l'incompréhension?
- Ces signaux diffèrent-ils selon le degré d'incompréhension.

Pour répondre à ces interrogations nous avons constitué un corpus audio-visuel à partir d'une expérience dans laquelle nous avons placé dix-sept sujets en situation d'incompréhension. Chaque sujet a été placé dans une situation avec interaction puis sans interaction. Des hachures ont été posées sur des endroits précis du discours afin d'estimer des degrés d'incompréhension.

Après annotation de notre corpus, nous avons procédé à un ensemble de tests statistiques à la suite desquels nous avons isolés quatre marqueurs auto-adressés :

- le marqueur "haussement de sourcils"
- le marqueur "froncement de sourcils"
- le marqueur "sourire inversé" présent lui aussi particulièrement en situation de forte incompréhension

Puis trois marqueurs à but communicatif :

- le marqueur "répétition partielle" présent particulièrement en situation de forte incompréhension

- le marqueur "question" présent particulièrement en situation d'incompréhension modérée
- le marqueur "supposition"
- le marqueur "sourire", particulièrement présent en situation de forte incompréhension

Nos analyses nous ont aussi permis de mettre en avant le fait que chaque sujet réagit aux perturbations dans le discours avec un délai régulier, ce qui nous montre que le sujet n'attend pas la fin de l'énoncé pour réagir.

Finalement, cette étude a su nous donner des résultats concluant par rapport à nos questions de départ. Il s'agit là d'un premier pas dans un large champ de recherche et ce travail offre des perspectives autour de l'incompréhension. La description de nos comportements pourraient être poussée afin de lever le voile sur certains aspects de nos marqueurs. De plus, une annotation plus poussée de notre corpus par plusieurs annotateurs pourraient corriger la possible trop grande subjectivité qu'il y a eu lors de cette tâche.

Au delà de cette recherche, les perspectives ne manquent pas. Nous pouvons en effet imaginer élargir le corpus que nous avons constitué à d'autres types d'incompréhension comme celle liée à un manque de connaissances communes. Nous pouvons élargir notre corpus à d'autres configurations d'interaction comme une conversation à plus de trois actants ou même un conférencier face à un auditoire.

Ensuite, la détection de ces marqueurs pourrait permettre des avancés notamment dans le domaine de la communication homme-machine pour ce qui est de l'augmentation de la fluidité dans le dialogue. Une application pouvant reconnaître l'incompréhension pourrait aussi s'intégrer dans le cadre de conférences afin de donner un indice au conférencier sur la compréhension de son discours par l'auditoire.

Bibliographie

Allwood, J., & Ahlsen, E. (1999). Learning how to manage communication, with special reference to the acquisition of linguistic feedback. *Journal of Pragmatics*, *31*(10), 1353-1389.

Boholm, M., & Allwood, J. (2010, May). Repeated head movements, their function and relation to speech. In *Proceedings of LREC workshop on multimodal corpora advances in capturing coding and analysing multimodality* (pp. 6-10).

Abric, J. C. (1996). *Psychologie de la communication: théories et méthodes*. A. Colin, chapitre 1.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on* (pp. 1-10). IEEE.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta analysis. *Personnel psychology*, *44*(1), 1-26.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication*, *52*(3), 566-580.

Bavelas, J., Gerwing, J., & Healing, S. (2017). Doing mutual understanding. Calibrating with micro-sequences in face-to-face dialogue. *Journal of Pragmatics*, *121*, 91-112.

Bavelas, J., Gerwing, J., & Healing, S. (2017). Doing mutual understanding. Calibrating with micro-sequences in face-to-face dialogue. *Journal of Pragmatics*, *121*, 91-112.

Bavelas, J. B., & Gerwing, J. (2011). The listener as addressee in face-to-face dialogue. *International Journal of Listening*, *25*(3), 178-198.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, *58*(2), 495-520.

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning: An integrated message model of language in face-to-face dialogue. *Journal of Language and social Psychology*, *19*(2), 163-194.

Bavelas, J. B., Chovil, N., Coates, L., & Roe, L. (1995). Gestures specialized for dialogue. *Personality and social psychology bulletin*, *21*(4), 394-405.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID-Corpus of Interactional Data-Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues*, *49*(3), pp-105.

Boholm, M., & Allwood, J. (2010, May). Repeated head movements, their function and relation to speech. In *Proceedings of LREC workshop on multimodal corpora advances in capturing coding and analysing multimodality* (pp. 6-10).

Carletta, J., Isard, A., Kowtko, J., & Doherty-Sneddon, G. (1996). *HCRC dialogue structure coding manual*. Human Communication Research Centre.

Clark, H. H., & Gerrig, R. J. (1990). Quotations as demonstrations. *Language*, 764-805.

Cerrato, L., & Skhiri, M. (2003). A method for the analysis and measurement of communicative head movements in human dialogues. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*.

Chovil, N. (1989) Communicative Functions of Facial Displays in Conversation. Unpublished Ph.D. dissertation. University of Victoria, Victoria, B.C.

De Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture*, *7*(1), 21-38.

de Fornel, M. (1992). «ALORS, TU ME VOIS?»: Objet technique et cadre interactionnel dans la pratique visiophonique, *Centre de recherche sur la culture technique*, Neuilly-sur-Seine

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*. Oxford University Press, USA.

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. Environmental psychology and nonverbal communication. *San Francisco: Human Sciences*.

Eberhard, K., & Nicholson, H. (2010, January). Coordination of understanding in face-to-face narrative dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern recognition*, *36*(1), 259-275.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, *9*(Aug), 1871-1874.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research*, *3*(2), 131-141.

Gardner, R. (1998). Between Speaking and Listening: The Vocalisation of Understandings1. *Applied linguistics*, *19*(2), 204-224.

Granström, B., & House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech communication*, *46*(3-4), 473-484.

Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human studies*, *9*(2-3), 205-217.

Graf, H. P., Cosatto, E., Strom, V., & Huang, F. J. (2002, May). Visual prosody: Facial movements accompanying speech. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on* (pp. 396-401). IEEE.

Ito, A., Wang, X., Suzuki, M., & Makino, S. (2005, November). Smile and laughter recognition using speech processing and face recognition from conversation video. In *Cyberworlds, 2005. International Conference on* (pp. 8-pp). IEEE.

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., & Bernstein, L. (2003, August). Optical phonetics and visual perception of lexical and phrasal stress in English. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)* (pp. 2071-2074).

Kießling, A., Kompe, R., Niemann, H., Nöth, E., & Batliner, A. (1993). " Roger"," Sorry"," I'm still listening": dialog guiding signals in information retrieval dialogs. In *ESCA Workshop on Prosody*.

McCarthy, M. (2003). Talking back:" Small" interactional response tokens in everyday conversation. *Research on language and social interaction*, *36*(1), 33-63.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Krych-Appelbaum, M., Law, J. B., Jones, D., Barnacz, A., Johnson, A., & Keenan, J. P. (2007). "I think I know what you mean": The role of theory of mind in collaborative communication. *Interaction Studies*, *8*(2), 267-280.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological science*, *15*(2), 133-137.

Mushin, I., Stirling, L., Fletcher, J., & Wales, R. (2000, October). Identifying prosodic indicators of dialogue structure: some methodological and theoretical considerations. In *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10*(pp. 36-45). Association for Computational Linguistics.

Navarretta, C., & Lis, M. (2013, October). Multimodal feedback expressions in Danish and Polish spontaneous conversations. In *NEALT Proceedings. Northern European Association for Language and Technology; 4th Nordic Symposium on Multimodal Communication; November 15-16; Gothenburg; Sweden* (No. 093, pp. 55-62). Linköping University Electronic Press.

Obin, N. (2012, September). Cries and whispers-classification of vocal effort in expressive speech. In *Interspeech*.

Özyürek, A. (2011). Language in our hands: The role of the body in language, cognition and communication, Radboud University

Scherer, K. (1999) "Appraisal Theory", *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power [Eds], pp. 637–663, John Wiley, New York.

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., ... & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech?. *Language and speech*, *41*(3-4), 443-492.

Tanaka, H., & Campbell, N. (2011, August). Acoustic features of four types of laughter in natural conversational speech. In *Proc. 17th International Congress of Phonetic Sciences (ICPhS), Hong Kong* (pp. 1958-1961).

Shannon, C. (1948). A Mathematical Theory of Communication-The Bell System Technical Journal, vol. 27, pag. 379-423, 623-656.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, *14*(1), 129-182.

Weigand, E. (1999). Misunderstanding: The standard case. *Journal of pragmatics*, *31*(6), 763-785.

Sigles et abréviations utilisés

AU : Action Unit (les unités d'action faciales développées par Paul Ekman)

AI : Phase avec Interaction

SI : Phase Sans Interaction

TAI : Phase Test Avec Interaction

TSI : Phase Test Sans Interaction

SP : Énoncé Sans Perturbation

P0 : Énoncé Perturbé en position 0

P1 : Énoncé Perturbé en position 1

P2 : Énoncé Perturbé en position 2

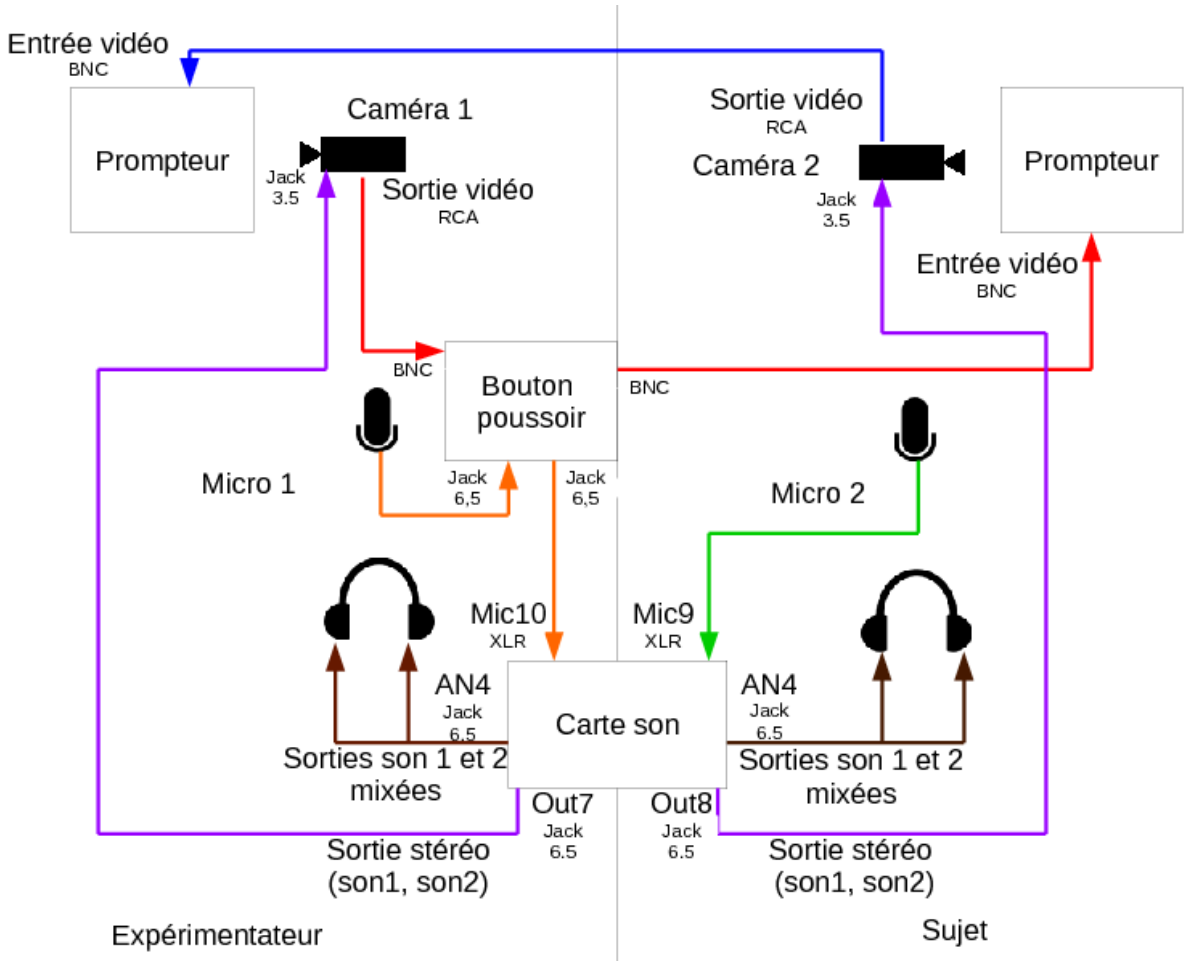
ET : Écart-type

Table des Figures

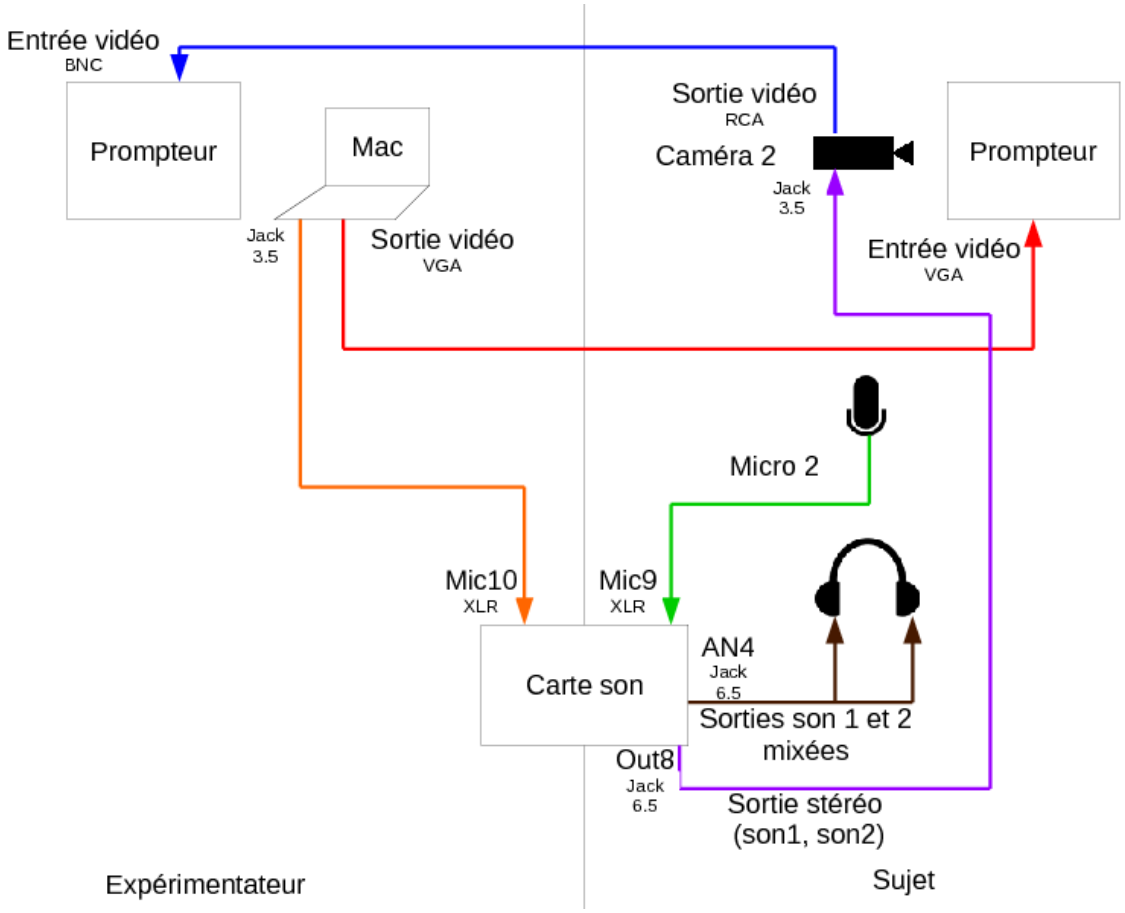
Table des Tableaux

Table des annexes

Annexe 1
Schéma d'installation de la phase avec interaction



Annexe 2
Schéma d'installation de la phase sans interaction



Annexe 3
Instructions de la carte test

tu passes à gauche de la robe

~~tu vas~~ au dessus du pied[0]

tu dois aller au dessus de la fraise

tu passes ~~au dessus~~ du lit[1]

tu dois aller au dessous ~~de la chaise~~[2]

tu vas à droite de l'os

~~tu passes~~ au dessus du verre[0]

tu passes au dessous du vase

tu dois aller à gauche ~~de la poire~~[2] *en passant par dessus*

tu vas ~~à droite~~ de la table[1]

tu passes au dessous ~~de la dent~~[2]

~~tu vas~~ à droite du chien[0]

Annexe 4
Instructions des cartes avec Interaction

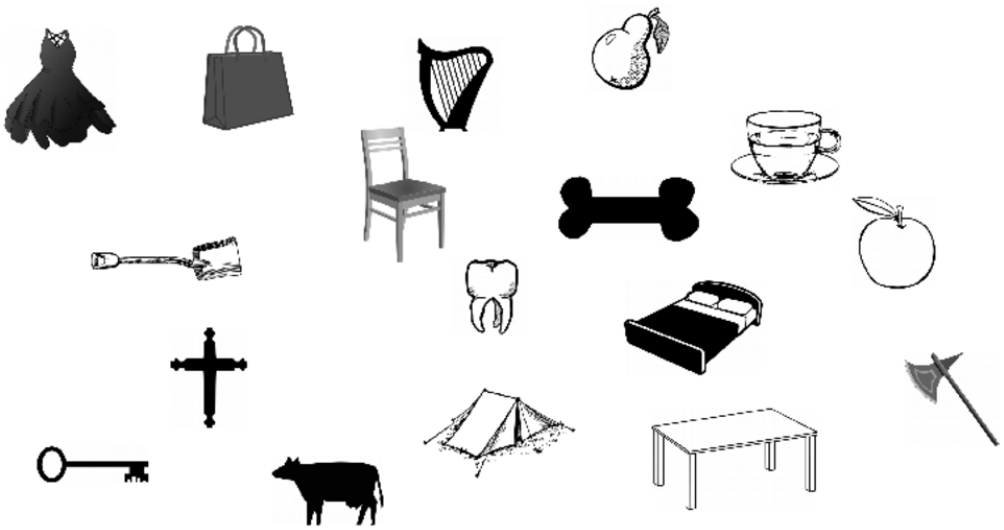
tu dois aller à droite ~~de la robe~~[2]	tu dois aller au dessous ~~de la feuille~~[2]
tu vas à gauche de la pelle	tu passes au dessus de la fraise
~~tu passes~~ au dessus de la croix[0]	tu vas à droite ~~de la poule~~[2] *en passant par dessus*
tu vas ~~au dessous~~ de la vache[1]	~~tu passes~~ à gauche de la hache[0]
tu passes à droite de la tente	tu vas ~~au dessus~~ de la poire[1] *par dessous la hache*
tu passes au dessous ~~de la dent~~ [2]	tu passes ~~à droite~~ de la flèche[1]
tu dois aller ~~à gauche~~ de la chaise[1]	tu vas au dessous de la table
tu vas au dessus ~~de la harpe~~[2]	tu passes à gauche ~~de la clé~~[2]
tu dois aller au dessous de la poire	~~tu vas~~ au dessous de la fleur[0]
~~tu dois aller~~ au dessus de la tasse[0]	~~tu dois aller~~ à droite de la flûte[0]
tu passes ~~à gauche~~ de la pomme[1]	tu dois aller à gauche de la robe
~~tu vas~~ à droite de la hache[0]	tu dois aller ~~au dessus~~ de chaise[1]
tu vas ~~à gauche~~ de la fleur[1]	tu passes ~~au dessus~~ de la pelle[1]
~~tu passes~~ à droite de la main[0]	tu dois aller à gauche ~~de la poire~~[2]
tu dois aller au dessus ~~de la jupe~~[2]	~~tu dois aller~~ au dessous de la hache[0]
tu passes au dessous de la table	tu passes à droite ~~de la dent~~[2]
tu dois aller ~~à droite~~ de la dent[1]	tu dois aller au dessus de la porte
tu vas au dessous ~~de la tente~~[2]	tu dois aller ~~au dessous~~ de la chaîne[1]
~~tu dois aller~~ à gauche de la fraise[0]	tu vas à droite de la chaise
tu vas au dessus de la croix	~~tu vas~~ à gauche de la flûte[0]
~~tu vas~~ au dessus de la porte[0]	tu passes au dessus ~~de la clé~~[2]
tu dois aller à droite de la harpe	tu passes à gauche de la tasse
tu vas à gauche ~~de la chèvre~~[2]	~~tu passes~~ au dessous de la pomme[0]
tu passes ~~au dessous~~ de la chaîne[1]	tu vas ~~à droite~~ de la feuille[1]

Annexe 5
Instructions des cartes sans interaction

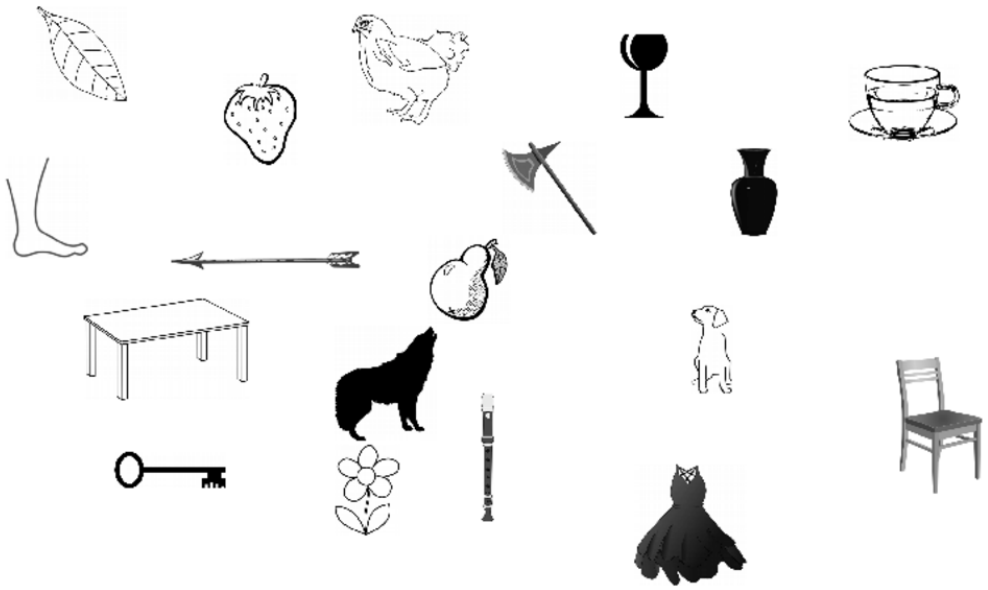
tu dois aller ~~au dessous~~ de la pomme[1]	tu vas au dessus de la fraise
tu passes à droite ~~de la vache~~[2]	tu passes ~~au dessous~~ de la poire[1]
tu dois aller à gauche ~~de la harpe~~[2]	tu dois aller ~~à droite~~ de la hache[1]
tu dois aller au dessus de la porte	tu vas ~~à gauche~~ de la pelle[1]
tu passes ~~au dessus~~ de la dent[1]	~~tu dois aller~~ à gauche de la table[0]
~~tu vas~~ à gauche de la pelle[0]	tu dois aller au dessus ~~de la feuille~~[2]
~~tu dois aller~~ au dessous de la tasse[0]	~~tu passes~~ à droite de la chèvre[0]
tu vas à droite de la flûte	tu passes au dessous de la robe
~~tu passes~~ au dessous de la croix[0]	tu vas au dessous ~~de la poule~~[2]
tu vas ~~à droite~~ de la flèche[1]	tu vas à gauche ~~de la pomme~~[2]
tu passes à gauche de la feuille *en passant par dessous la flèche*	tu dois aller à droite de la chaîne
tu passes au dessus ~~de la fraise~~[2] *en passant par dessous la feuille*	~~tu vas~~ au dessus de la flèche[0]
tu passes ~~à droite~~ de la porte[1]	tu dois aller à droite ~~de la main~~[2]
tu passes à gauche ~~de la chaîne~~[2]	tu vas à gauche de la harpe
tu vas ~~au dessus~~ de la tente[1]	tu passes au dessous ~~de la fraise~~ [2]
tu passes au dessus de la harpe	tu vas au dessus ~~de la pelle~~[2]
tu dois aller au dessous ~~de la poule~~[2]	tu vas ~~au dessous~~ de la flûte[1]
tu vas au dessous de la main	tu dois aller ~~à gauche~~ de la fleur[1]
tu vas à droite ~~de la fleur~~[2]	~~tu passes~~ au dessus de la pomme[0]
~~tu passes~~ à gauche de la jupe[0]	tu passes à droite de la tasse *en passant au dessus*
tu dois aller ~~au dessus~~ de table[1]	~~tu vas~~ à droite de la clé[0]
tu dois aller à gauche de la poire	tu passes ~~à gauche~~ de la poire[1]
~~tu dois aller~~ à droite de la croix[0]	~~tu dois aller~~ au dessus de la flèche[0]
~~tu vas~~ au dessous de la robe[0]	tu dois aller au dessous de la chaise

Annexe 6
Cartes Avec Interaction

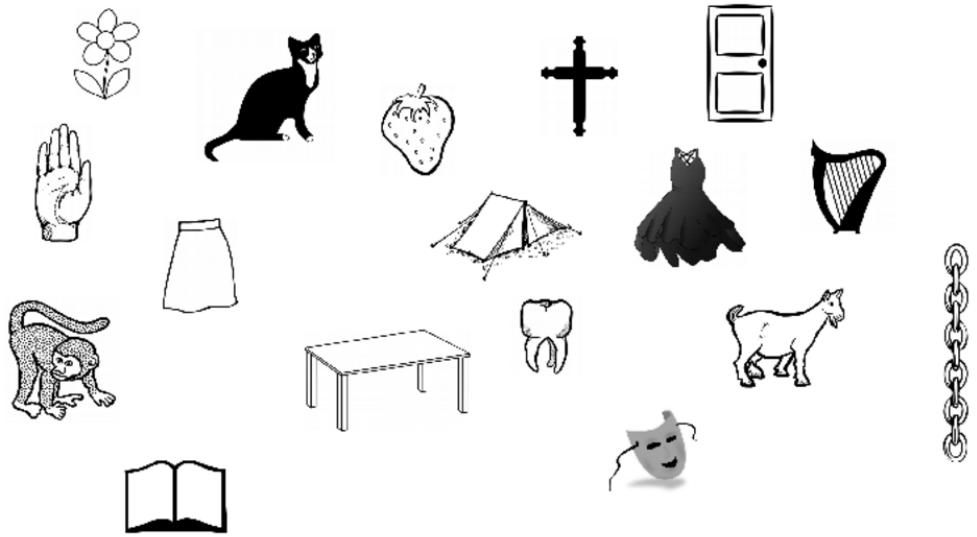
START



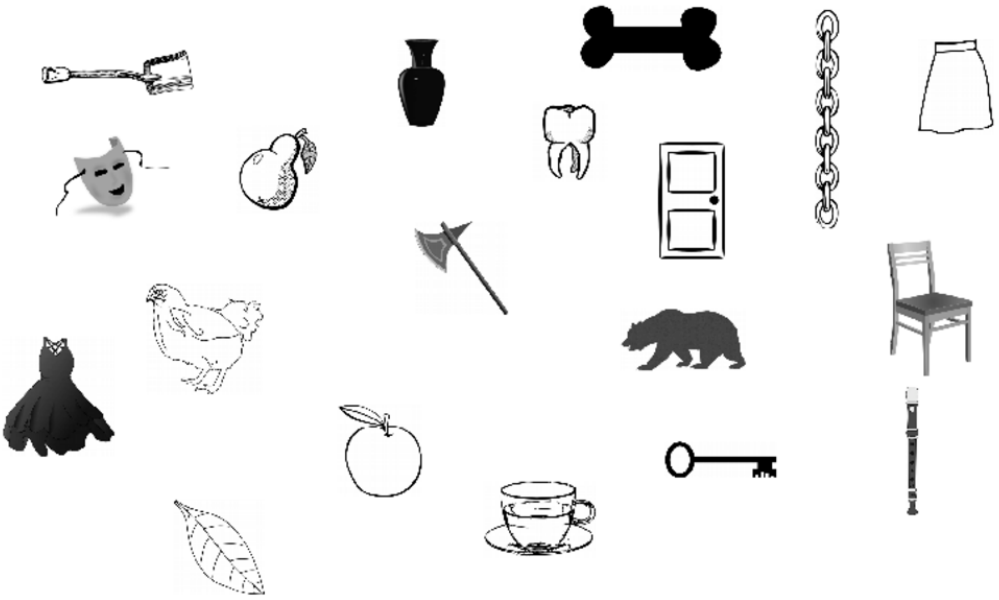
START



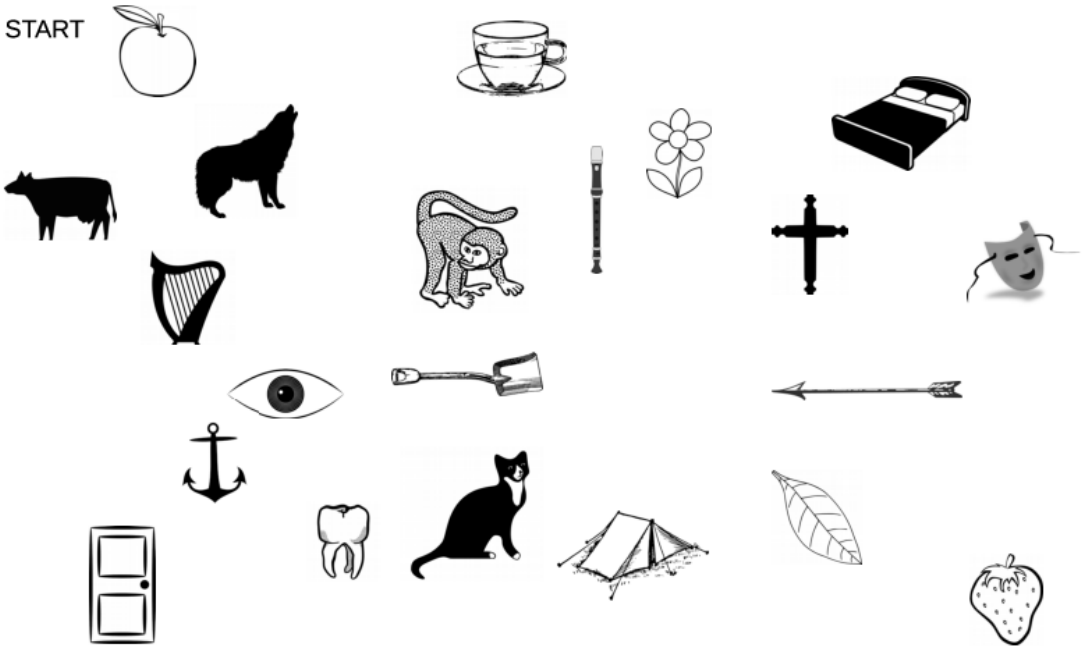
START



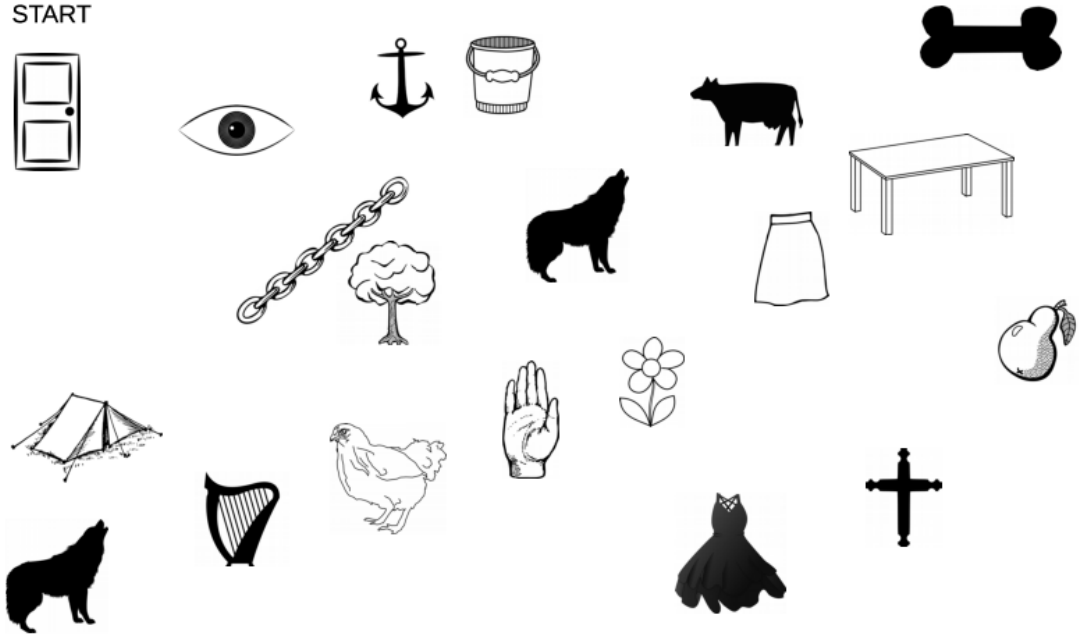
START



Annexe 7
Cartes Sans Interaction



START



START

Table des matières

MOTS-CLÉS : Incompréhension, non-verbal, dialogue, rétroaction

RÉSUMÉ

Ce mémoire porte sur les signaux verbaux et non verbaux qu'émet le récepteur d'un message lors d'un dialogue face à face lorsque qu'il est placé en situation d'incompréhension. Cette recherche se concentre sur deux aspects : Quels sont parmi les signaux émis, ceux auto-adressés et ceux à but communicatif et est ce que le degré d'incompréhension a une influence sur le type de signal produit. Cette recherche a mené à la constitution d'un corpus audio-visuel annoté avec les comportements de quinze sujets. Des analyses statistiques effectuées sur ces annotations ont permis d'isoler neuf comportements relatifs à l'incompréhension. À partir des comportements faciaux isolés nous avons pu extraire un ensemble de paramètres grâce à un outil de détection de mouvements faciaux pour identifier des paramètres pertinents dans l'optique de développer un système de détection.

KEYWORDS : Incomprehension, nonverbal, dialogue, back-channel

ABSTRACT

This thesis addresses the verbal and nonverbal signals that the listener of a message produces during a face to face dialogue in an incomprehension situation. This research focuses on two aspects : identifying among these behaviors, which that are self-adressed and those with a communicative purpose and does the incomprehension degree influence the type of behavior? For this research we built an audio-visual corpus annotated with the behaviors of fourteen subjects. A statistical analysis of these annotations led to singling out nine behaviors related to incomprehension. From these behaviors, we extracted parameters with a facial behavior analysis toolkit in order to check our results from these low-level descriptors.