Modélisation de dialogues pour personnage virtuel narrateur

Ovidiu Serban, Anne Bersoult, Zacharie Alès, Élise Lebertois, Emilie Chanoni, François Rioult, Alexandre Pauchet

Modélisation de dialogues pour personnage virtuel narrateur

Ovidiu Şerban [1], Anne Bersoult [3], Zacharie Ales [1], Elise Lebertois [3], Emilie Chanoni [3], François Rioult [2], Alexandre Pauchet [1]

1. *INSA Rouen - LITIS*
   Avenue de l'Université - BP 8
   76801 Saint-Etienne-du-Rouvray Cedex, France
   {prénom.nom}@insa-rouen.fr
2. *Université de Caen - Greyc*
   Campus Côte de Nacre, F-14032 Caen Cedex, France
   francois.rioult@unicaen.fr
3. *Université de Rouen - Psy-NCA*
   Rue Lavoisier 76821 Mont Saint-Aignan Cedex , France
   {prénom.nom}@univ-rouen.fr

RÉSUMÉ. *Dans l'optique de concevoir un Agent Conversationnel Animé narratif et affectif, cet article montre l'importance de l'interaction dans le processus de narration à des enfants. Deux méthodes d'extraction de régularités ont été appliquées à des dialogues de narration parent-enfant afin de les modéliser. Les modèles extraits ont été évalués au cours d'une session de narration interactive de type Magicien d'Oz durant laquelle les enfants ont communiqué avec un personnage virtuel et un adulte en visio-conférence. L'expérience montre que les enfants s'engagent bien dans l'interaction avec le personnage virtuel mais que les modalités pour interagir diffèrent légèrement de celles utilisés avec l'adulte.*

ABSTRACT. *To develop a narrative affective virtual agent, this article shows the importance of interaction modalities in storytelling environments. Building a narrative dialogue model requires a state transition that is computed automatically using various pattern extraction methods. These models are tested in a "Wizard-of-Oz" context, in a storytelling environment where the children are challenged to interact with a virtual character, by modelling several key points of interaction. We compare the level of engagement in two different narration contexts: virtual character or adult in video conference mode. Our experiments showed that the engagement exists, but the modality of the interaction feedback varies in the two contexts.*

MOTS-CLÉS : *Modélisation du dialogue, Motifs dialogiques, ACA narrateur, Magicien d'Oz*

KEYWORDS: *Dialogue Modelling, Dialogue Pattern Extraction, Narrative ECA, Wizard of Oz*

1. Introduction

Depuis quelques années, de nouvelles interfaces homme-machine se développent sous la forme d'Agents Conversationnels Animés (ACA) (Cassell *et al.*, 2000). Avec l'émergence des environnements numériques, et plus particulièrement de systèmes de narration participative, les situations d'interactions enfant-ACA sont de plus en plus fréquentes. Le dialogue avec un ACA, en remplacement du temps passé devant la télévision par exemple, devrait être bénéfique au développement social, langagier et cognitif des enfants, comme certains travaux tendent déjà à le montrer (Ryokai *et al.*, 2003). En effet, une situation de narration interactive avec un ACA rassemble l'intérêt des enfants pour le monde virtuel et l'avantage d'un moment d'interaction à l'origine de l'apprentissage de compétences socio-cognitives.

Cependant, afin d'être accepté par les enfants, un ACA dédié à la narration interactive doit adopter un comportement correspondant aux habitudes d'interaction adulte-enfant afin de faciliter la compréhension de ce dernier : le modèle dialogique de l'agent doit être adapté aux compétences sociales et langagières de l'enfant. La conception d'un modèle de dialogue pour un ACA, qu'il soit spécifique à la narration interactive ou non, est une tâche difficile impliquant de nombreux mécanismes : traitement de signaux multimodaux (parole, gestes, regards, etc.), reconnaissance et génération de langage naturel, gestion du dialogue, modélisation des émotions, prosodie et comportement non verbal. En particulier, la gestion de la multi-modalité et des émotions dans le dialogue reste à ce jour insuffisante au sein des ACA, bien que ces aspects soient essentiels pour des interactions efficaces (Swartout *et al.*, 2006).

Dans le cas particulier des enfants, ces derniers utilisent certaines de leurs compétences dans l'interaction avec ce nouveau partenaire qu'ils semblent considérer comme un interlocuteur acceptable. Quelques études montrent que la communication fonctionne bien entre eux (Oviatt, 2000 ; Ryokai *et al.*, 2003). Cependant, bien que certaines spécificités d'interaction enfant-agent virtuel aient été mises en évidence, notamment au niveau de la fluence du discours et des modalités d'expression (Oviatt, 2000 ; Buisine, Martin, 2003), très peu d'éléments existent pour caractériser les interactions d'un enfant avec ses partenaires non-humains. Il est ainsi difficile d'affirmer que la simple intégration d'un modèle de dialogue narratif adulte-enfant puisse être intégré dans un ACA sans problème d'acceptabilité. En particulier, le phénomène d'attente généré par un personnage virtuel très expressif, connu sous le nom de "vallée dérangeante" (*the uncanny valley*) (Mori, 1970), est peu évaluable chez l'enfant.

L'objectif de cet article est double. Il s'agit tout d'abord de proposer une modélisation du dialogue permettant d'extraire de manière semi-automatique un modèle d'interaction dialogique à partir de dialogues réels. Pour ce faire nous proposons, à partir d'un corpus, d'une part d'extraire des motifs dialogiques, et d'autre part de prédire les interactions de l'interlocuteur afin de guider le dialogue. Dans un second temps, cette modélisation est appliquée à un ensemble de dialogues de narration entre parents et enfants, afin d'en extraire un modèle de dialogue narratif. Le modèle extrait est évalué en session de narration interactive, joué par un humain en visio-conférence et par un

avatar en situation de Magicien d'Oz, afin d'estimer l'impact de l'incarnation auprès des enfants. Notre projet vise ainsi à construire un environnement familier, centré autour de l'activité de narration. La population visée est un ensemble d'enfants entre 5 et 10 ans, c'est-à-dire ayant déjà acquis une théorie de l'esprit (Astington, Baird, 2005), ayant déjà eu des contacts avec les nouvelles technologies.

Un état de l'art sur le modélisation du dialogue pour ACA et la méthodologie Magicien d'Oz est présenté section 2. La méthode proposée pour modéliser des dialogues narratifs ainsi que le corpus de dialogues étudiés sont décrits section 3. Les détails des procédures d'extraction de régularité dans des dialogues sont décrits section 4 et le modèle extrait est expliqué section 5. La section 6 est consacrée à une plate-forme originale permettant de faciliter la mise en place d'expérimentations de type Magicien d'Oz. Cette plate-forme a permis de tester les modèles de dialogue narratif extraits lors d'une expérimentation dont les résultats sont fournis section 7. Enfin, cet article se termine par une courte conclusion ainsi qu'un exposé de nos futurs travaux.

2. Dialogue, ACA et Magicien d'Oz

2.1. Modèles du dialogue pour ACA

Les ACA sont des interfaces autonomes et anthropomorphiques, incarnés par des personnages animés aux compétences multi-modales : langage naturel, expressions du visage, regards, attitudes et gestes (Cassell *et al.*, 2000). Les ACA peuvent être catégorisés selon leur expressivité : systèmes de type présentateurs, interactions face à face (un ACA avec un humain) et conversations multipartites (plusieurs ACA et utilisateurs) (André, Pelachaud, 2010). Les projets récents sur les ACA se focalisent sur l'interactivité en perfectionnant les expressions faciales et le comportement non verbal afin d'améliorer la qualité générale de l'agent (Cassell *et al.*, 2000 ; Pelachaud, 2009). Greta (Pelachaud, 2009), MARC (Courgeon *et al.*, 2009) et le projet européen SEMAINE (Schröder, 2010) sont de bons exemples des capacités actuelles des ACA.

Dans le domaine de la narration, les agents virtuels intelligents, qu'ils soient incarnés ou non, peuvent être utilisés en tant que personnages expressifs (ex : (Seif El-Nasr, Wei, 2008)) et en tant que narrateurs. Le projet GV-LEx (Gesture and voice for an expressive reading), par exemple, a pour but de fournir aux robots Nao (Gouaillier *et al.*, 2009) et à l'ACA Greta (Pelachaud, 2009) la capacité de lire du texte sans ennuyer l'auditeur (Gelin *et al.*, 2010). Ce projet propose d'utiliser une intonation expressive ainsi que des gestes tout en parlant afin de produire des narrations crédibles.

En ce qui concerne les systèmes et modèles du dialogue pouvant être intégrés dans les ACA, plusieurs approches existent.

– **L'approche *à états finis*** (voir par exemple (McTear, 2004)) représente la structure du dialogue par un automate à états finis dans lequel chaque énoncé conduit à un nouvel état. En pratique, cette approche est limitée aux systèmes de dialogue directifs.

– **L'approche *par formulaire (frame-based)*** représente le dialogue comme un processus de remplissage de formulaire contenant des entrées prédéfinies (voir par

exemple (Aust *et al.*, 1995)). Les contributions possibles sont fixées à l'avance.

– **L'approche *par planification*** (exemple : (Allen, Perrault, 1980)) combine la reconnaissance de plans et la théorie des Actes de Langage (Searle, 1969). Cette approche est complexe du point de vue calculatoire et requiert des composants très avancés de TAL afin d'inférer les intentions du locuteur.

– **L'approche *logique*** représente le dialogue et son contexte par un formalisme logique et utilise des mécanismes tels que l'inférence et les jeux de dialogue (voir par exemple (Hulstijn, 2000)). La plupart des travaux concernant les approches logiques ne sont actuellement qu'au stade de la théorie.

– **Les approches *par apprentissage*** proposent des techniques telles que l'apprentissage par renforcement (Frampton, Lemon, 2009) afin de modéliser le dialogue via des processus de Markov. Ces approches requièrent un gros travail d'annotation.

En raison de la complexité des systèmes de dialogue complet, la plupart des ACA existants n'intègrent que des processus basiques de gestion du dialogue, tels que la détection de mots-clés au sein d'une approche de type frame-based ou à états finis. Companions (Cavazza *et al.*, 2010) est plus qu'un simple ACA, puisqu'il s'agit d'un compagnon engagé dans une interaction à plus long terme avec l'utilisateur. Le scénario de Companions est centré autour de la thématique *"How was your day ?"* afin de permettre un dialogue relativement ouvert. Les différents énoncés sont traités par un extracteur d'informations qui reconnaît des événements clefs comme les promotions individuelles, les rendez-vous, les relations entre collègues... Les interactions sont contrôlées par un modèle du dialogue à base de règles. Semaine (Schröder, 2010) est un interlocuteur affectif dont le personnage virtuel perçoit l'émotion de l'utilisateur au travers d'un ensemble d'éléments multi-modaux et réagit selon cette émotion. Le système de dialogue sous-jacent est une simple reconnaissance de mots clefs. Le Virtual Human Toolkit (Hartholt *et al.*, 2013) est une plate-forme générique permettant de développer des ACA sous une architecture par composants. Une bibliothèque de composants est fournie, couvrant l'ensemble des fonctionnalités nécessaire à un système interactif : reconnaissance de la parole et transcription automatique, gestion du dialogue, génération de comportement non verbal pour un ACA... La gestion du dialogue est effectuée par modèle construit par apprentissage sur des corpus de type question-réponse générique. En résumé, la gestion du dialogue est limitée dans les ACA actuels (Swartout *et al.*, 2006), mais efficace pour certains types de dialogue.

2.2. Magicien d'Oz

Afin d'évaluer le modèle de dialogue d'un ACA indépendamment des autres composants le constituant, une conception itérative peut être utilisée. Ainsi, de nombreux groupes de recherche abordent ce problème d'abord par un scénario de type "Magicien d'Oz" afin, soit de constituer un corpus initial, soit d'évaluer uniquement le modèle de dialogue. Ce scénario peut être utilisé itérativement pour compléter le modèle initial.

Le paradigme du "Magicien d'Oz" est largement utilisé pour les évaluations des interfaces multimodales (Aubergé *et al.*, 2003). C'est une méthode qui permet de re-

cueillir des interactions obtenues dans des conditions écologiques et identiques pour les sujets et dont le contenu est contrôlé. Ce paradigme consiste à mettre en interaction un utilisateur naïf avec une machine. La personne pense communiquer avec un système machine complexe et autonome, alors qu'il s'agit d'un partenaire humain qui pilote et contrôle le système en amont (magicien), donnant ainsi à l'utilisateur l'illusion d'interagir avec un système doué d'intelligence artificielle.

La plupart des expériences réalisées avec un Magicien d'Oz ne sont pas réutilisables (Salber, Coutaz, 1993 ; Douglas-Cowie *et al.*, 2008 ; Otto *et al.*, 2011) de par leur lien étroit avec le contexte considéré. DiaWOZ-II propose une interface textuelle dans un contexte de tutorat. (Whittaker *et al.*, 2002) se servent d'une interface web pour simuler des scénarios de dialogues dans un restaurant. (Munteanu, Boldea, 2000) se basent également sur une interface textuelle simple mais intègrent un modèle du dialogue à états finis que doit suivre le pilote. De nouveaux états peuvent être ajoutés en temps réel au modèle si nécessaire. Plus récemment les expérimentation de type Magicien d'Oz se dotent d'interfaces multimodales. SUEDE (Klemmer *et al.*, 2000) propose reconnaissance et synthèse de discours simulés tout en conservant la gestion du dialogue par le magicien. Artur (Bälter *et al.*, 2005) introduit un ensemble d'images décrivant un contexte d'apprentissage. Le projet SEMAINE a lui aussi débuté avec une expérience en Magicien d'Oz (McKeown *et al.*, 2010), qui a mené à un modèle d'interaction simple utilisé actuellement dans la version finale de SEMAINE.

2.3. Interactions avec un personnage virtuel / robot

Du point de vue de la personnification, (Cassell, 2000) a introduit le concept d'interaction en face-à-face avec un avatar animé. Le niveaux de détails utilisés pour représenter un personnage virtuel souvent très élevés induisent des attentes particulières de la part de l'interlocuteur quant aux capacités interactives d'un ACA. Cependant, les faibles compétences conversationnelles et les réactions non naturelles de l'agent déçoivent et mènent à des interactions peu naturelles entre l'humain et le système. Ce phénomène est appelé la "vallée dérangeante" (*the uncanny valley*) (Mori, 1970). Ce type de comportements affectent l'empathie des utilisateurs envers les agents (Beale, Creed, 2009). Pour surmonter ces inconvénients, l'agent doit répondre à la frustration des utilisateurs (Klein *et al.*, 2002), devenir plus empathique (Ochs *et al.*, 2008 ; Prendinger, Ishizuka, 2005), émotionnel (Poggi *et al.*, 2005) et réagir au moment approprié avec une posture ou un geste adapté à la situation (Prepin, Pelachaud, 2013).

L'influence des personnages virtuellement animés (conversationnels ou non) sur la perception humaine est appelée "persona effect". Des études pédagogiques (Mourdridou, Virvou, 2002) et utilisant les jeux sérieux avec des enfants (Prendinger *et al.*, 2003) ont montré l'existence d'un lien entre la présence d'un personnage virtuel et les performances de l'utilisateur. Cependant, (Miksatko *et al.*, 2010) concluent qu'un tel impact n'existe pas. (Grynszpan *et al.*, 2008) ont réalisé une étude multi-modale, via un Magicien d'Oz, qui a révélé une forte influence de l'agent sur les performances de patients atteints d'autisme de haut niveau.

Dans le cadre d'interactions avec des enfants, l'influence des personnages virtuels n'a pas été profondément étudiée. Une expérience réalisée par (Oviatt, 2000) révèle que les enfants âgés de 6 à 10 ans génèrent moins d'irrégularités rompant la fluidité de leur discours (ou *diffluences*) lorsqu'ils parlent à une méduse animée plutôt qu'à un adulte. De plus, les enfants sont très attirés par ce nouvel interlocuteur et acceptent aisément de s'impliquer dans le dialogue. (Ryokai *et al.*, 2003) ont réalisé une étude sur l'utilisation d'un Agent Conversationnel Animé (ACA), appelé Sam, dans un scénario de tutorat pour enfants. L'objectif était d'accélérer le processus d'apprentissage de la lecture et de l'écriture à travers la narration. Dans ce contexte, Sam raconte une histoire dans un environnement collaboratif. Le personnage virtuel ressemble à un ami d'école mais raconte des histoires permettant à l'enfant d'acquérir des compétences nécessaires à l'alphabétisation. Cette étude à démontré que l'implication sociale efficace de l'enfant avec l'ACA a permis un apprentissage rapide de nouveaux mots ou de constructions linguistiques complexes.

Des études similaires existent également avec des robots. Le même niveau d'engagement est observé chez les enfants autistes (Kozima *et al.*, 2005 ; Robins *et al.*, 2005), en situation de tutorat (Han *et al.*, 2005) ou dans des situations de développement cognitif précoce (Von Hofsten, Rosander, 2007). Le potentiel de ces deux types d'interlocuteurs, pour des applications en éducation et en narration, est similaire. L'étude d'une approche conversationnelle ou narrative avec un robot étant légèrement plus complexe et coûteuse, les ACA sont généralement favorisés pour ces types d'applications.

2.4. Approche proposée

Jusqu'alors, très peu de projets ont permis de mettre en place des environnements de narration utilisant un ACA (Gelin *et al.*, 2010). Les travaux présentés dans la suite de ce document proposent d'utiliser un personnage virtuel afin de susciter l'engagement d'un enfant lors d'une tâche de narration interactive.

Concernant le modèle de dialogue, toutes les approches présentées précédemment utilisent comme représentation des structures de données régulières, extraites manuellement ou apprises automatiquement à partir d'un corpus de dialogues ou de traces. Ces structures de données ne permettent de représenter que des motifs d'interactions linéaires. Cependant la gestion du dialogue implique plusieurs dimensions et non une seule (Bunt, 2011), comme par exemple les dimensions de la tâche, de la gestion des tours de parole ou encore la gestion des obligations sociales. Le modèle d'interaction d'un ACA nécessite la gestion de tous les aspects associés aux interactions humaines (gestion de tâches individuelles et collectives, feedbacks, aspects affectifs, obligations sociales, etc.), exprimés suivant différentes modalités (sémantique, prosodie, gestes, expressions, etc.). Nous proposons d'extraire un modèle de dialogue multidimensionnel à partir d'un corpus de dialogues réels de narration entre un parent et son enfant. Il s'agit de combiner planification au service de la résolution de la tâche - prédiction et planification des interventions de l'enfant - et une gestion plus réactive par motifs dialogiques. Une représentation matricielle encodant le dialogue devrait

permettre de tenir compte du caractère multidimensionnel du dialogue. Le modèle de dialogue construit sera alors évalué au moyen d'un Magicien d'Oz.

3. Modélisation du dialogue pour la narration interactive

La méthode de modélisation du dialogue proposée est présentée Figure 1 :

– Une **collecte et numérisation** d'un corpus de dialogues au format audio ou vidéo est effectuée. Le corpus que nous considérons dans cet article est composé d'histoires enfantines racontées par des parents à leur enfant ;

– l'étape **transcription et codage** consiste à produire des données brutes à divers niveaux de détails (tours de parole, énoncés, onomatopées, pauses, etc.) selon les caractéristiques que l'on souhaite exhiber ;

– une phase d'**extraction de connaissances**, suivant un schéma de codage spécifique, est ensuite appliquée aux énoncés encodés afin d'obtenir une description précise des comportements dialogiques. Les dialogues sont alors considérés comme annotés ;

– une phase d'**extraction de régularités** est appliquée aux annotations. Les régularités extraites constituent le modèle.

– le modèle peut alors être **exploité**.

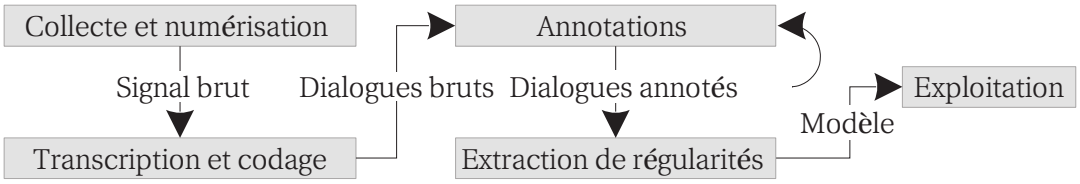

Figure 1. Méthodologie d'analyse du dialogue.

Dans les sous-sections suivantes, nous présentons l'expérimentation réalisée pour collecter le corpus de dialogues de narration ainsi que le schéma de codage utilisé pour obtenir une représentation matricielle des dialogues.

3.1. Corpus de dialogues de narration d'histoires enfantines

La narration d'histoires d'un parent à son enfant est une situation classique participant au développement des enfants. Les contextes sociaux et langagiers apportés par l'adulte sont nécessaires à l'enfant dans son processus d'apprentissage des compétences socio-communicatives, cognitives et morales. Les enfants développent une *théorie de l'esprit* (Astington, Baird, 2005) durant leurs premières années et deviennent ainsi capables d'assimiler le fait qu'une personne est déterminée par ses propres intentions, émotions et états mentaux. Ce développement n'est possible qu'au travers des situations sociales de dialogue. Le discours des adultes concernant les états mentaux se révèle être un médiateur d'apprentissage du concept de cognition sociale - grâce à une participation active au dialogue et à des interactions dynamiques.

Dans cette étude, nous utilisons un corpus de 90 dialogues entre parents et enfants âgés de 3, 4 et 5 ans, filmés en situation de récit d'histoires enfantines (10 enfants

par tranche d'âge × 3 histoires différentes). Ces enregistrements sont retranscrits et annotés par un unique annotateur, suivant une *grille mentaliste* (Chanoni, 2009), afin de faire ressortir les informations relatives aux états mentaux (croyances, volition, émotions, etc.) contenues dans les énoncés. La longueur moyenne des dialogues est de 89,3 énoncés. Seules les expressions verbales sont retranscrites.

3.2. Représentation des dialogues

Comme le souligne Bunt, la gestion du dialogue est multi-niveaux (Bunt, 2011). Afin de concevoir un modèle du dialogue multi-dimensionnel, les annotations sont représentées matriciellement. Chaque énoncé est caractérisé par un vecteur d'annotations dont les composantes correspondent aux différentes dimensions de codage : une ligne par énoncé et une colonne par espace/dimension de codage.

Le Tableau 1 présente un exemple de dialogue provenant du corpus collecté. Chaque énoncé est caractérisé par un numéro de ligne, un locuteur (P : parent, E : enfant), une transcription et des annotations encodées suivant 5 dimensions :

– la première colonne caractérise la nature de l'énoncé : une (A)ffirmation, une (Q)uestion, une demande d'attention, générale (G) ou concernant l'histoire (D) ;

– la seconde colonne définit la référence de l'énoncé : à un personnage (P), à l'auditeur (H) ou au narrateur (R) ;

– la troisième colonne est dédiée aux états mentaux. Les interlocuteurs peuvent exprimer une (E)motion, une (V)olition, une cognition observable (B) ou non (N), une déclaration épistémique (K), une hypothèse (Y) ou une (S)urprise. La surprise se distingue des autres émotions de par son lien avec les croyances ;

– les deux dernières colonnes représentent les explications par (C)ause/conséquence, (O)pposition ou empathie (M), qui peuvent être utilisées, soit pour expliquer l'histoire (J), soit pour préciser une situation par l'évocation d'un contexte personnel (F).

Par exemple, la ligne 35 est encodée ainsi : l'énoncé est une affirmation (A) portant sur un état mental se référant à un personnage - "*Babar*" - (P) ; l'état mental correspondant - "*sait*" - se réfère à une cognition non observable (N) ; "*Mais*" dénote une justification par opposition (O) ; enfin, l'énoncé se réfère à l'histoire (J).

La construction de cette représentation matricielle nécessite un processus d'annotation manuel, semi-automatique et/ou automatique - un pour chaque dimension/colonne. Les matrices obtenues peuvent être vues comme des séquences de vecteurs d'annotations. La représentation fournie ici correspond à un schéma de codage particulier, mais doit bien évidemment être adaptée en fonction des besoins applicatifs.

4. Extraction de régularités dans des dialogues annotés

Nous proposons deux approches d'extraction de régularités : un calcul de similarité par programmation dynamique permettant de collecter des motifs dialogiques

Tableau 1. Représentation matricielle des annotations d'un dialogue de narration d'une histoire enfantine entre un parent et son enfant

Ligne	Locuteur	Enoncé	Annotations				
25	Parent	T'inquiète pas	A	P	E	-	-
26	Parent	Donc là ils se cachent	A	P	B	-	-
27	Parent	Ils cherchent	A	-	-	-	-
28	Parent	qui pourrait avoir pris la couronne.	Q	-	-	-	-
29	Enfant	Elle est dedans, elle est dedans la couronne.	A	-	-	-	-
30	Parent	Donc là ils suspectent plein de monde, Cornélius, Céleste, la vieille dame...	A	P	Y	C	J
31	Parent	Qui a bien pu prendre la couronne ?	Q	-	-	-	-
32	Enfant	La couronne elle est dedans.	A	-	-	-	-
33	Parent	Tu crois ?!	Q	H	K	-	-
34	Enfant	Oui.	A	-	-	-	-
35	Parent	Mais Babar il ne sait pas qu'elle est dedans.	A	P	N	O	J
36	Parent	Donc il se dit que c'est une bombe, la couronne	A	P	N	C	J
37	Parent	ou je ne sais quoi.	A	R	N	-	-

et une méthode de prédiction d'événements se concentrant sur la caractérisation des interactions de l'enfant.

4.1. Extraction de motifs dialogiques

Avec notre représentation matricielle, un *motif dialogique* est défini comme un ensemble d'annotations dont la disposition apparaît - de manière exacte ou approchée - dans plusieurs dialogues. Un motif peut contenir des annotations non adjacentes en ligne ou en colonne (i.e. un motif peut avoir des trous), et deux instances d'un même motif peuvent contenir des insertions, des suppressions ou des substitutions. Deux motifs sont donc considérés comme similaires si leur distance d'édition est faible.

La figure 2 présente la méthode utilisée pour extraire un ensemble de motifs dialogiques pertinents. Elle est composée d'une extraction de régularités basée sur un alignement de matrices par programmation dynamique, permettant de collecter un ensemble de paires de motifs similaires, suivi d'une étape de clustering afin de regrouper les motifs dialogiques récurrents. Le processus de clustering est appliqué à un graphe de similarités calculé durant l'alignement de matrices.

La méthode d'extraction de motifs en deux dimensions s'apparente à l'alignement de matrices. Il s'agit d'une généralisation de la distance d'édition locale entre deux vecteurs de caractères. La distance d'édition *ed* (ou distance de Levenshtein) entre deux vecteurs de caractères s_1 et s_2 correspond au coût minimal des trois opérations d'édition élémentaires (insertion et suppression de caractères, ainsi que la substitution d'un caractère par un autre) permettant de transformer s_1 en s_2. Un alignement local de deux matrices de caractères s_1 et s_2, de tailles respectives $m_1 \times n_1$ et $m_2 \times n_2$,

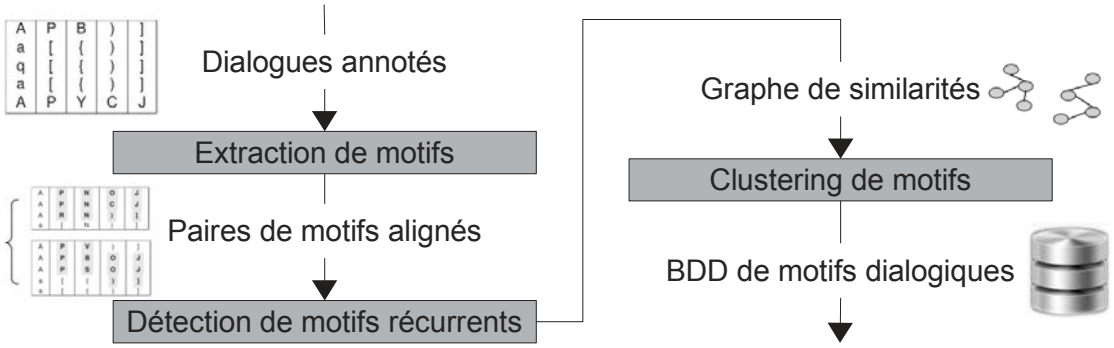

Figure 2. Extraction de motifs dialogiques.

consiste à chercher les portions de s_1 et s_2 qui sont les plus similaires (parmi toutes les portions de s_1 et s_2). Pour ce faire, une table à 4 dimensions T de taille $(m_1+1) \times (n_1+1) \times (m_2+1) \times (n_2+1)$ est calculée, de telle sorte que $T[i][j][k][l]$ soit égal à la distance d'édition locale entre $S_1[0..i-1][0..j-1]$ et $S_2[0..k-1][0..l-1]$, $\forall i \in [\![1, m_1-1]\!]$, $j \in [\![1, n_1-1]\!]$, $k \in [\![1, m_2-1]\!]$ et $l \in [\![1, n_2-1]\!]$. Dans notre méthode, le calcul de T est obtenu par minimisation d'une formule de récurrence. Une fois T calculée, le meilleur alignement local est obtenu en effectuant un algorithme de tracé arrière à partir de la position où T atteint sa valeur maximale. Ce tracé arrière permet d'inférer les caractères faisant partie de l'alignement. La figure 3, commentée en section 5, présente un exemple d'alignement issu de notre corpus. Pour de plus amples informations sur l'extraction de motifs en 2D, se référer à (Lecroq *et al.*, 2012).

L'alignement de matrices permet d'extraire les motifs par paires. Nous les regroupons à l'aide de différents algorithmes de clustering (voir tableau 2). L'idée sous-jacente est que les clusters les plus conséquents représentent les comportements les plus communs, tandis que les petits clusters reflètent des comportements plus marginaux. Une matrice de similarité entre les différents motifs est calculée grâce à une distance d'édition globale appliquée aux paires de motifs détectés. Cette matrice est utilisée comme entrée des algorithmes de clustering.

Tableau 2. Indice de Dunn en fonction du nombre de clusters pour les heuristiques implémentées. Le caractère '-' est utilisé lorsqu'une solution n'est pas produite pour un nombre de clusters donné. Les valeurs surlignées correspondent aux meilleur(s) résultat(s) pour chaque colonne.

Méthode	**Référence**	**Nombre de clusters trouvés**					
		5	**20**	**50**	**80**	**116**	**150**
Single-Link	(Florek *et al.*, 1951)	41	97	183	270	320	360
CHAMELEON	(Karypis *et al.*, 1999)	458	605	**628**	-	-	-
ROCK	(Guha *et al.*, 2000)	520	600	621	626	629	630
Spectral Clustering (SC)	(Von Luxburg, 2007)	277	**658**	563	155	194	226
SC selon Shi et Malik	(Von Luxburg, 2007)	524	615	**628**	**631**	631	**632**
SC selon Jordan et Weiss	(Von Luxburg, 2007)	**555**	616	**628**	630	631	**632**
Propagation d'affinité	(Frey, Dueck, 2007)	-	-	-	-	**632**	-

Cette méthode a été testée sur le corpus de dialogues de narration. Durant la phase d'extraction, 1740 motifs dialogiques ont été collectés, de taille variant entre 10 et 124 énoncés pour une moyenne de 28,9.

Le nombre de solutions des méthodes de clustering étant trop élevé pour une comparaison au cas par cas, l'indice de Dunn (Dunn, 1973) a été utilisé afin d'évaluer les méthodes. Si, pour une solution donnée, s_{ij} représente la similarité entre deux motifs i et j, et $c(i)$ désigne l'indice du cluster contenant i, l'indice de Dunn est définit par

$$\frac{\min_{c(i)\neq c(j)} s_{ij}}{\max_{c(k)=c(l)} s_{kl}}.$$

Ainsi, les solutions comportant un indice de Dunn élevé sont susceptibles d'être pertinentes, car composées de clusters compacts et séparés. Le nombre de clusters étant inconnu, les méthodes ont été testées sur plusieurs valeurs. Le tableau 2 illustre une partie représentative des résultats de l'indice de Dunn. Les meilleures méthodes semblent être la propagation d'affinité et les méthodes de type spectral clustering.

Les motifs dialogiques pouvant être extraits sont représentatifs des conventions dialogiques. Ils permettent ainsi d'avoir un modèle réactif pour répondre à des sollicitations non prévues lors d'une tâche. Dans le cadre de la narration interactive enfantine, il s'agit à la fois d'une représentation des mécanismes utilisés par les parents pour expliquer des situations à leur enfant, mais surtout leur comportement dialogique en cas de réaction de l'enfant.

4.2. Prédiction des interventions de l'enfant

Cette section est consacrée à la définition d'un modèle du dialogue permettant de stimuler l'interaction. Dans cette optique, les interventions de l'enfant doivent être finement modélisées en se concentrant sur la *prédiction d'événement*. Nous recherchons des séquences d'événements dialogiques entraînant une interaction particulière afin de s'en servir comme plan pour générer les interventions d'un ACA narrateur.

Les contributions sur l'extraction de connaissances à partir de séquences sont principalement consacrées aux épisodes et à la classification de séquences. La prédiction d'événement sur des données discrètes n'y est que très peu abordée (Antunes, Oliveira, 2001). Nous proposons de découper les données en *tours de parole*, caractérisés par un ensemble d'énoncés successifs provenant d'une seule personne (ici le parent ou l'enfant). Le problème revient à prévoir la fin du tour. Dans ce but, nous considérons des séquences de séries de vecteurs d'annotations (une série de vecteurs d'annotations par tour de parole) se terminant par une intervention de l'enfant. Par exemple, les séquences du tableau 1 sont :

$$< (APE)(APB)(A)(Q) >, < (APYCJ)(Q) >, < (QHK) >, ...$$

Pour extraire les régularités menant à la fin des séquences, les épisodes sont explorés par projections récursives grâce à un algorithme glouton. Dans l'exemple ci-

dessus, l'algorithme débute avec l'épisode $< (Q) >$, commun à toutes les fins de séquences. L'algorithme est ensuite appelé une nouvelle fois sur les séquences projetées $< (APE)(APB)(A) >, < (APYCJ) >$. (A) est ajouté à l'épisode qui devient $< (A)(Q) >$, qui est lui même projeté une nouvelle fois : les séquences résultantes sont $< (APE)(APB) >$, ... L'explosion combinatoire est limitée par deux contraintes antimonotones : la fréquence d'apparition et la distance moyenne en nombre d'énoncés à la fin de la séquence.

Au cours du traitement - dans lequel la séquence est parcourue de la fin vers le début - les épisodes obtenus ne sont pas nécessairement tous appropriés à la prédiction de la fin du tour de parole. Supposons, par exemple, que chaque séquence commence et termine par une (Q)uestion, l'algorithme décrit précédemment donnera comme prédicteur de fin $< (Q) >$, bien qu'il soit aussi un bon prédicteur de début. Pour éviter ce type de résultats défavorables, la distance moyenne de chaque épisode au début de la séquence doit être prise en compte. Si cette dernière est trop faible, l'épisode n'est pas conservé. Ce processus assure que les régularités extraites sont pertinentes.

Le processus d'extraction fournit un très grand nombre d'épisodes. Afin que l'expert puisse manuellement les évaluer, il est nécessaire d'en limiter le nombre. Dans cette optique, une approche par clustering de trajectoire (Lee *et al.*, 2007) a été adoptée, en considérant que les épisodes sont des séquences de déplacements entre deux ensembles de vecteurs d'annotations. Les déplacements sont classés et un représentant est obtenu pour chaque classe. Ce regroupement permet de passer de plusieurs centaines d'épisodes à seulement quelques dizaines de représentants.

Ainsi, un ensemble de séquences fréquemment utilisés par les parents et efficaces pour générer une réaction des enfants peut ainsi être collecté et utilisé comme trame de base pour un scénario de narration.

5. Analyse des modèles obtenus

L'évaluation des modèles calculés montre qu'un agent narrateur doit être interactif avec l'enfant (au travers de questions et de demandes d'attention) et ce d'autant plus avec les enfants en bas âge. Il apparaît essentiel de solliciter les enfants afin qu'ils interagissent. De plus, la compréhension des émotions et états mentaux des personnages peut être améliorée par une explication du comportement entraîné par l'état mental.

Dans la suite de cette section, l'évaluation des modèles réalisée par une psychologue spécialiste des interactions parents-enfant est détaillée. Ces modèles devraient permettre d'expliquer les comportements relatifs aux états mentaux observés et, à terme, d'être intégrés dans un ACA narrateur afin d'en guider le comportement.

5.1. Motifs dialogiques

L'extraction de motifs dialogiques a permis de collecter un ensemble de motifs et de les regrouper selon leur score de similarité calculé par programmation dynamique.

La figure 3 présente un exemple d'alignement de motifs. Chacune des instances de ce motif n'a été identifiée qu'une seule fois de manière exacte, mais se retrouve plusieurs fois de manière approchée, au sein d'un même cluster de motifs dialogiques. Ce motif montre que les parents parlent, tout d'abord, des causes ou des conséquences du comportement du personnage (P, C, J) sans référence à l'état mental. Après quelques affirmations ou questions, les parents insistent sur la justification du comportement du personnage (ligne 6), puis le mettent en relation directe avec l'état mental du personnage (ligne 7). Enfin, le parent vérifie que l'enfant a compris en posant des questions ou en demandant son attention (ligne 8).

Ligne	**Locuteur**	**Dialogue B3 (4 ans)**					**Locuteur**	**Dialogue C8 (5 ans)**				
0	Parent	A	-	-	-	-	Parent	A	-	-	-	-
1	Parent	A	**P**	E	**C**	**J**	Parent	A	**P**	-	**C**	**J**
2	Parent	Q	**-**	-	**-**	**-**	Enfant	A	**-**	-	**-**	**-**
3	Enfant	A	**-**	-	**-**	**-**	Parent	A	**-**	-	**-**	**-**
4	Parent	A	**-**	-	**-**	**-**	Parent	A	**-**	-	**-**	**-**
5	Parent	A	**-**	-	**-**	**-**	Parent	A	**-**	-	**-**	**-**
6	Parent	A	**P**	-	**C**	**J**	Parent	A	**P**	-	**C**	**J**
7	Parent	A	**P**	**E**	**-**	**-**	Enfant	Q	**P**	**E**	**-**	**-**
8	Parent	**Q**	**-**	**-**	**-**	**-**	Parent	**D**	**-**	**-**	**-**	**-**
9	Enfant	A	-	-	-	-	Parent	A	P	E	-	-

Figure 3. Exemple d'alignement de deux motifs dialogiques.

Ce motif démontre parfaitement qu'il n'est pas suffisant de nommer un état mental pour l'expliquer. En effet, le développement narratif implique une démonstration pratique de la théorie de l'esprit. Le motif décrit le lien entre le comportement du personnage et l'état mental, le second expliquant le premier.

5.2. Prédiction d'interaction

Nous décrivons ici les conditions nécessaires à une amélioration significative de la narration interactive, en fonction de l'âge de l'enfant (3, 4 ou 5 ans). Le tableau 4 résume les modèles des interactions de l'enfant (voir section 3.2). Pour chaque âge, les modèles sont caractérisés par :

– leur *longueur moyenne*, qui correspond au nombre moyen d'énoncés entre le modèle et l'interaction de l'enfant. Plus une séquence est courte, moins il y a d'énoncés entre elle et l'intervention de l'enfant ;

– le *modèle*, qui décrit une séquence d'annotations. Par exemple la séquence E-Q symbolise une annotation E suivie, plus ou moins tard, d'une annotation Q. Les annotations peuvent ne pas être dans la même dimension ;

– la *fréquence*, qui est le pourcentage de fois où le modèle apparaît.

Les demandes d'attention (codées D, par exemple "regarde !" ou "tu as vu ?"), essentielles pour la narration interactive, sont présentes pour tous les âges. Plus l'enfant

3 ans		
longueur	**modèle**	**fréquence**
3,2	**E-Q**	**10,4%**
3,4	D-Q	16,8%
3,5	J-Q	9,6%
3,5	D-Q-Q	9,6%
3,5	E-J	8,8%
4,3	D-E	12,8%
4,3	D-J	8,0%
5,4	B	10,4%
5,6	V	13,6%

4 ans		
longueur	**modèle**	**fréquence**
2,1	D-Q	14,9%
2.2	**E-Q**	**7,5%**
2,2	Q-Q	12,7%
2,6	D-E	10,4%
2,8	D-D	11,2%
3,5	J	14,9%
3,8	B	7,5%
4,0	E-E	7,5%
4,1	E-D	6,7%
4,3	V	6,7%

5 ans		
longueur	**modèle**	**fréquence**
1,9	Q	35.4%
2,2	**E-E**	**9,1%**
2,6	J-D	8,1%
2,7	E-D	6,1%
3,1	J-E	8,1%
3,4	V	13,1%
3,7	D-E	7,1%
3,8	J-J	6,1%
4,1	E-J	7,1%

Figure 4. Longueurs moyennes et fréquences de séquences en fonction de l'âge.

est âgé, plus sa réaction à la demande d'attention est rapide. Plus l'enfant est jeune, plus les demandes d'attention doivent être répétées ou ponctuées de questions. Ceci peut s'observer dans des séquences telles que D-D ou D-Q ou D-Q-Q.

110 séquences contenant une demande d'attention ont été recensées dans les dialogues des enfants de 3 ans, 58 pour les enfants de 4 ans. Les enfants de 3 ans interagissent en effet après un nombre d'énoncés moyen compris en entre 3,5 et 4,6. Les enfants de quatre ans réagissent plus rapidement (entre 2,2 et 2,8 énoncés) : l'efficacité du modèle (l'inverse de la longueur moyenne des séquences) s'améliore avec l'âge. Par contre, les parents se comportent différemment avec les enfants de 5 ans : les demandes d'attention sont moins fréquentes et soit associées à des états mentaux (D-E ou E-D) soit à des justifications (D-C ou C-D). Nous n'avons dénombré que 21 séquences comportant des demandes d'attention, rapidement suivies d'une interaction de l'enfant (entre 2,6 et 2,7 énoncés). Les séquences comprenant des justifications (codées J, par exemple "puis, Leo casse le château !") sont essentielles au processus d'interaction émotionnelle narrative.

En conclusion, nous souhaitons mettre l'accent sur certains points notables :

– les séquences contenant des justifications sont fréquemment associées à divers indices (émotion, demande d'attention ou question). Dans ce contexte, l'interaction de l'enfant ne survient qu'entre 3,1 et 4,3 énoncés après le modèle ;

– la longueur des interactions décroît avec l'âge, de 3,2 à 1,9 énoncés ;

– le nombre d'énoncés auxquels sont associées des émotions est quasiment équivalent pour tous les âges. Néanmoins, plus l'enfant est âgé, plus les séquences d'émotions sont variées. Les séquences complexes (émotions et justifications : J-E ou E-J) n'apparaissent qu'avec les enfants les plus âgés ;

– à l'exception des demandes d'attention, les modèles les plus efficaces (en rouge et gras dans le tableau 4) contiennent toujours des émotions (E-Q ou E-E).

5.3. Motifs et prédiction comme modèle de dialogue

À l'aide des motifs dialogiques représentatifs extraits, il est possible de construire un scénario de narration utilisant des mécanismes explicatifs similaires à ceux utilisés par les parents. En d'autres termes, une suite de motifs dialogiques correctement choisi permet de raconter une histoire. Afin de susciter des réactions particulières, il est également possible d'introduire dans ce scénario linéaire un ensemble de points clefs, comme par exemple des erreurs d'interaction. Ces étapes peuvent se construire à partir des séquences extraites du modèle prédictif.

De la même manière, en cas de réaction non suscitée et non prévue, le modèle de dialogue doit intégrer les motifs simples permettant de ramener l'attention de l'enfant vers la tâche. Nous sélectionnons ici volontairement des motifs courts, permettant un retour à la narration en un tour de parole.

6. Evaluation du modèle de dialogue narratif par Magicien d'Oz

Le modèle de dialogue narratif précédemment extrait a vocation à être intégré au sein d'un ACA pour des séances de narrations interactives d'histoires enfantines. Une des difficultés importantes pour la recherche en interaction homme-machine est de collecter des résultats issus d'une situation expérimentale pour laquelle un seul élément est évalué. Afin d'évaluer notre modèle de dialogue, il n'est pas possible de l'intégrer au sein d'un ACA compte tenu des performances actuelles des différents éléments le composant (transcription, détection d'expressions faciales, reconnaissances de postures, etc.). Afin de nous affranchir de ces difficultés, il a été choisi d'évaluer l'engagement des enfants en situation d'interaction narrative, en comparant ses réactions en face d'un adulte en visio-conférence, avec ses réactions face à un personnage animé en situation de Magicien d'Oz. Dans les deux situations, notre modèle est utilisé pour guider la narration et les interactions avec les enfants. Il s'agira ainsi d'évaluer si l'enfant s'ajuste de façon spécifique en fonction de la nature de son interlocuteur.

Plusieurs aspects techniques doivent être garantis par le projet :

1. La collecte de données multimodales : une nouvelle plate-forme est développée afin de soutenir l'infrastructure de la collecte de données et qui permet une interaction riche entre l'enfant et une configuration de type Magicien d'Oz.

2. L'élaboration de scénario : il s'agit de choisir et d'intégrer une histoire spécifique adaptée au niveau langagier des enfants et de formaliser l'ensemble du dispositif expérimental. Compte-tenu des résultats obtenus confirmant la littérature sur l'acquisition de la théorie de l'esprit (Astington, Baird, 2005), la population visée doit être constituée d'enfants âgés de plus de 5 ans.

3. Le modèle de gestion du dialogue doit être cohérent avec le modèle précédemment extrait pour les enfants de 5 ans, correspondant aux habitudes d'interaction narrative parent-enfant.

4. Afin d'engager davantage l'enfant dans l'interaction, nous proposons une série de points clés dans l'histoire où plus d'attention est nécessaire. Nous appelons ces points: *les erreurs d'interaction*. Ils sont liés au modèle de prédiction des interactions.

6.1. Scénario suscitant l'interaction

Afin de susciter un engagement de l'enfant dans l'interaction, le point essentiel est de motiver celle-ci par des séquences d'actions dialogiques menant à une réaction de l'enfant (utilisation du modèle de prédiction des interventions de l'enfant). Le scénario doit donc comporter des séquences où l'enfant est impliqué dans une interaction et ses réactions doivent être prises en compte. Par ailleurs, afin que l'illusion d'interaction soit optimale, des séquences d'interactions hors scénario doivent également être prévues afin de ramener le dialogue vers la tâche de narration interactive. Toutes ces séquences doivent être en adéquation avec les motifs dialogiques correspondant aux standards d'interaction adulte-enfant en situation de narration interactive.

De manière à standardiser notre expérimentation, un scénario, accompagné d'images, a été élaboré à partir d'une histoire («Le Ballon Perché») dont la durée de narration est de 10 minutes. Le contexte de narration nous garantit un déroulement et un contenu contrôlé. Pour autant le contexte de narration entre un enfant et un adulte peut s'avérer assez pauvre en interaction si ce dernier ne pose aucune question. Cette interaction peut être suscitée si nécessaire, à l'aide de demandes d'attention ou de questions portant sur les états mentaux, comme cela a été observé en situation.

Par ailleurs, de la même façon qu'Aubergé (2003) utilise des perturbations dans le comportement attendu pour susciter les états internes émotionnels aux sujets, nous proposons un scénario de narration dans lequel nous avons glissés des erreurs interactionnelles de façon à stimuler les réactions de l'enfant envers le narrateur.

Les échanges ont ainsi été encouragés de deux manières. Premièrement, le narrateur pose des questions à des moments précis du scénario à l'enfant concernant le ou les personnages ou l'histoire suivant la forme des motifs extraits. Deuxièmement, des erreurs interactionnelles ont été intégrées au scénario et sont destinées à perturber, interrompre ou altérer la communication. Ces perturbations permettent de susciter des

réactions de la part des enfants. Si l'enfant se sent dans une situation d'interaction, alors une réparation ou un ajustement de la part de l'enfant est nécessaire pour que celle-ci perdure. De fait, si l'enfant réagit, c'est qu'il se considère en interaction. Le second avantage est que nous obtenons ainsi des réactions authentiques dont le déclenchement est contrôlé. Les données sont comparables et nous permettent de différencier la nature de réaction de l'enfant en fonction du narrateur avec lequel il interagit.

Les six erreurs interactionnelles sont disposées dans le scénario comme indiqué sur la Figure 5, de manière à pouvoir comparer les réactions des enfants en fonction de la nature du narrateur. Ces 6 erreurs sont de 3 types différents :

– 2 erreurs de compréhension (C), par exemple une erreur sémantique (le narrateur dit le mot "*carotte*" au lieu du mot "*botte*" attendu)

– 2 erreurs relatives aux émotions (E), par exemple une erreur d'interprétation de l'émotion du personnage (le narrateur dit "*on dirait qu'il chante !*" alors que le personnage est en fait en colère.)

– 2 erreurs d'attention (A) qui altèrent le partage d'informations auditives ou visuelles entre le narrateur et l'enfant. Par exemple l'écran devient subitement noir pendant le narrateur continu de décrire l'image et pose des questions à l'enfant.

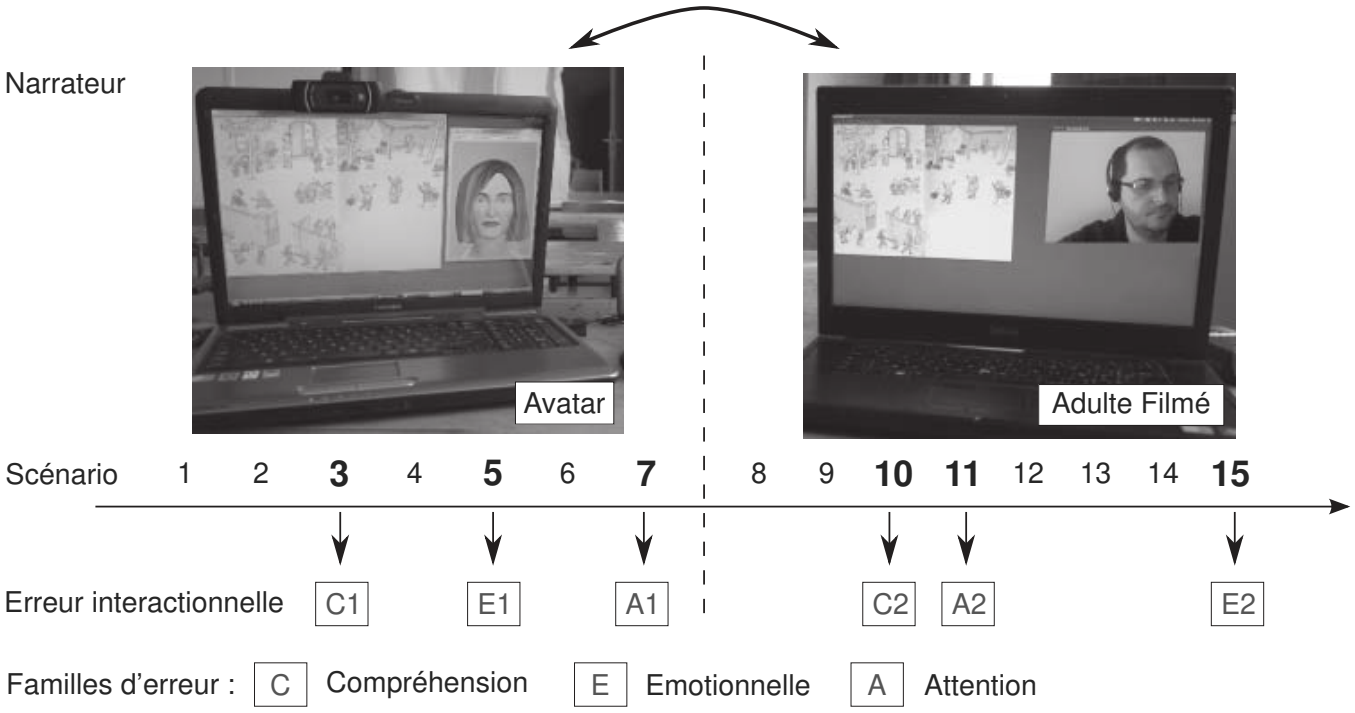

Figure 5. Scénario composé de 15 Images et de 6 erreurs interactionnelles.

L'histoire est séparée en 2 parties équivalentes, chacune d'entre elle est narrée soit par l'avatar (Magicien d'Oz), soit par l'adulte filmé (visio-conférence). L'ordre de narration (Avatar puis Adulte filmé ou Adulte filmé puis Avatar) est contrebalancé. À gauche de la fenêtre du narrateur, l'histoire est présentée grâce à 15 diapositives d'illustration qui apportent un support à l'histoire (Figure 5).

Le pointage est un comportement indispensable lors de la narration entre un enfant et un adulte en situation écologique et permet une attention conjointe indispensable à l'interaction (Scaife, Bruner, 1975). Un pointeur a donc été intégré au système de

manière à désigner les éléments décrits lors du récit permettant de s'assurer qu'une attention conjointe et précise est possible entre le narrateur et l'enfant.

En supplément du scénario, et toujours dans cet esprit d'augmenter le naturel et la flexibilité de l'interaction avec l'enfant, un certain nombre de petites phrases interactives, hors scénario, ont également été enregistrées. Il s'agit d'un ensemble de phrases qui ne sont pas directement liés au contexte de l'histoire, tels que : "*OK*", "*tu as raison*", "*On continue ?*" et qui ont comme objectif de permettre au pilote de répondre au discours et aux questions spontanées de l'enfant. Enfin, pour que celui-ci comprenne bien l'histoire racontée, il est fondamental d'utiliser des gestes de pointage. Une représentation d'un déictique a donc été intégrée au système de manière à ce que l'expérimentateur puisse désigner les éléments voulus au cours du récit.

Avant le début de l'histoire, une première prise de contact permet de familiariser le sujet avec le dispositif et l'interlocuteur, de s'installer au mieux dans l'interaction. Cette phase d'habituation est nécessaire pour que l'enfant comprenne bien que le contexte de l'expérience autorise une véritable interaction. Nous entamons la situation avec un moment de questions-réponses concernant le nom, l'âge, la classe de l'enfant et le narrateur se présentant lui aussi en fonction des questions de l'enfant (par exemple, ACA : "*tu as quel âge ?*", E : "*8 ans et demi !*", ACA : "*haa ! Tu es un grand maintenant !*", E : "*et toi ?*", ACA : "*moi, j'ai 20 ans !*").

L'histoire raconte l'aventure d'un garçon de l'école qui a décidé de jouer au ballon avec ses amis pendant la récréation. Malencontreusement le ballon est envoyé sur un toit. Les amis tentent de récupérer le ballon et décident tour à tour de lancer une chaussure, un cartable, etc. qui restent eux aussi coincés sur le toit. Ils sont alors invités à entrer en classe alors qu'ils n'ont pas pu récupérer leurs affaires. En classe, les enfants redoublent d'ingéniosité pour que le maître ne s'aperçoive pas qu'il manque une chaussure à l'un et qu'un autre n'a pas son cartable. Finalement, une énorme tempête arrive et fait tomber tous les objets du toit.

À la fin de l'expérience, pour chaque enfant, des questions au sujet de ces erreurs interactionnelles sont incluses dans une enquête finale. En outre, ils sont invités à résumer l'histoire et interroger sur les détails de toutes les "choses bizarres" (i.e. erreurs interactionnelles) qui se sont produites pendant l'expérience.

En comparaison aux travaux antérieurs, notre étude emploie donc un scénario formalisé et le "magicien" doit veiller à son exécution. Afin de simplifier sa tâche, nous avons conçu une plate-forme appelée le Toolkit d'Annotation En ligne (*Online Annotation Kit* - OAK) qui couvre toutes ces exigences.

6.2. OAK : une Plate-forme de Magicien d'Oz

OAK unifie différentes plates-formes et concepts en un seul outil. OAK est générique et suffisamment simple pour être utilisé pour la collecte de données en temps réel. Il n'exige que des compétences de manipulation simples. La conduite de l'avatar a été simplifiée au point où toutes les actions sont très intuitives. Un autre point

clé de la plate-forme est l'annotation en ligne, supportée par l'activation de chaque point clé du scénario à un instant donné. Cela donne une idée de l'ordre d'exécution des actions, des durées de celles-ci, et permet de formaliser une trace du dialogue. OAK a été utilisé en expérimentations réelles avec un bon niveau de satisfaction de la part du pilote et des enfants. En outre, la collecte de données annotées est simplifiée, puisqu'elle est effectuée de manière automatique tout au long de l'expérimentation.

6.2.1. Formalisation du scénario

Un scénario dans OAK est représenté par un automates à états finis, sous la forme d'un ensemble d'états et d'observations. Les états sont des actions qui sont exécutées par le moteur ou traduits directement en langage BML [1]. Les observations correspondent à des éléments de la perception du monde réel, formalisés sous forme de notes dans le scénario OAK et à partir desquelles le déroulement de l'expérimentation est construit. Alors que l'enchaînement des états est indispensable puisqu'il fixe les différentes actions communicatives, les observations sont optionnelles et servent uniquement à expliciter le scénario. Pour garantir la traçabilité du déroulement de l'expérimentation, toutes les transitions sont enregistrées et une même observation peut relier deux états. Cette caractéristique rend le rend scénario déterministe.

Afin d'illustrer le caractère générique d'OAK, nous décrivons la formalisation d'un scénario d'interaction différent de celui utilisé pour la narration collaborative d'histoires enfantines. La figure 6 présente cet exemple de scénario à trois états (s1-s3), dont les transitions sont annotés par des observations (o1-o5).

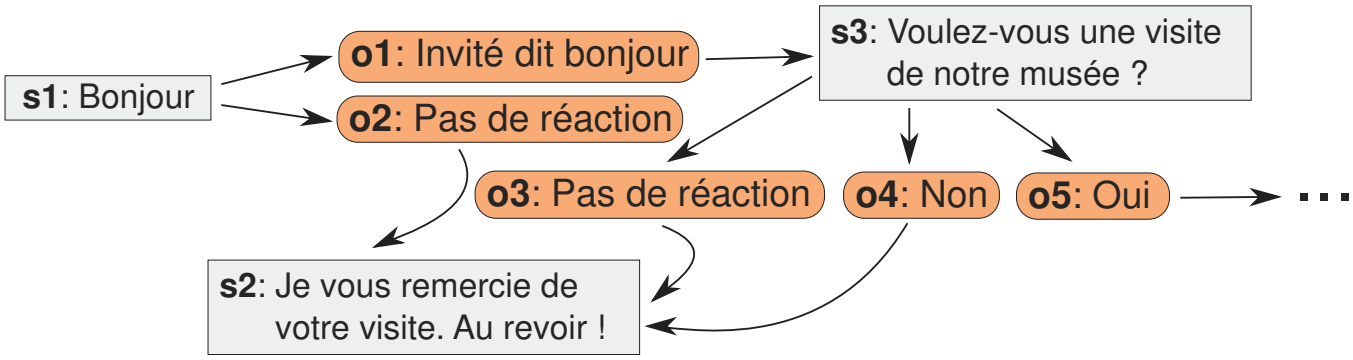

Figure 6. Exemple de scénario formalisé avec OAK.

6.2.2. Interfaces de pilotage et du sujet

La Figure 7 est la vue du pilote. Elle présente l'ensemble du scénario, c'est-à-dire la collection de tous les états possibles. Un état peut être exécuté à tout moment, autant de fois que nécessaire. La bibliothèque de "petites phrases interactives", pouvant être utilisées en dehors de la trame principale du scénario, se trouve à droite. En haut de cette bibliothèque, un menu permet de sélectionner le mode de visualisation: aucun (pas de visualisation), vidéo et avatar. Le mode Aucun correspond au début de

1. Le BML (*Behaviour Markup Language*) est un langage XML décrivant le comportement verbal et non-verbal spécifique pour un agent virtuel animé humanoïde (Kopp *et al.*, 2006).

l'expérience, avec la phase de description de l'installation. Les deux autres modes correspondent aux deux étapes du scénario de narration interactive.

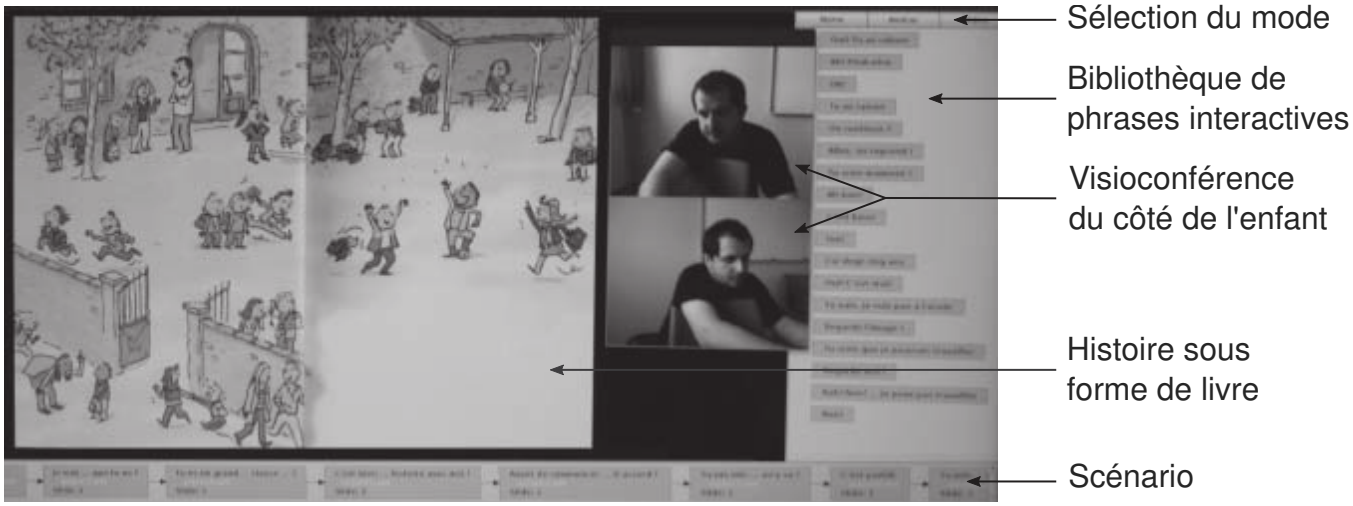

Figure 7. Vue du poste du pilote.

La Figure 8 est une vue de l'interface du côté de l'enfant. L'enfant a devant lui deux interfaces : l'une pour la partie de l'histoire narrée par l'ACA et une seconde pour la partie de l'histoire narrée par l'adulte en visio-conférence. Sur la gauche, le narrateur est Poppy (un des quatre personnages animés de la plate-forme SEMAINE (Schröder, 2010)), alors qu'il s'agit sur la droite d'un flux vidéo. Poppy a été utilisée afin d'avoir un narrateur féminin en phase de magicien d'Oz tout comme en phase de visio-conférence. Cette configuration nécessite deux webcams, ce qui nous a permis de filmer l'enfant à partir d'angles différent. Le flux vidéo récupéré est aussi envoyé à la vue de pilote. L'installation de la vidéo conférence est effectuée en utilisant plusieurs canaux de communication, construite avec Gstreamer pour Linux (Taymans *et al.*, 2001). Toutes les vidéos récupérées sont enregistrées en copies multiples, pour assurer la sauvegarde. Les illustrations de l'histoire se trouvent à gauche des narrateurs. Les images sont numérisées et synchronisées entre les 3 vues. En outre, le pilote peut utiliser la souris pour pointer sur des aspects importants de l'histoire.

Tous les composants de OAK sont entièrement personnalisables, grâce à des fichiers de configuration indépendants en XML. Les actions sont converties BML (Kopp *et al.*, 2006) par un interpréteur d'actions et envoyées au personnage animé.

Au cours de cette expérience, la narration ainsi que les phrases interactives sont fixées par le scénario. Le prototype permet au magicien de pointer à la souris certaines étapes du scénario. L'animation de la tête de l'ACA est prédéfinie à l'avance en fonction du contexte de l'histoire. Afin de limiter toute dérive, une bibliothèque de phrases interactives permettant de réorienter l'utilisateur vers le scénario principal, a été ajoutée (voir Figure 7).

En utilisant les données récupérées lors de l'expérimentation, nous sommes en mesure de procéder à une analyse statistique basée sur plusieurs descripteurs d'interaction. L'objectif est de comparer les réactions de l'enfant en fonction du type de narrateur qu'il a en face de lui.

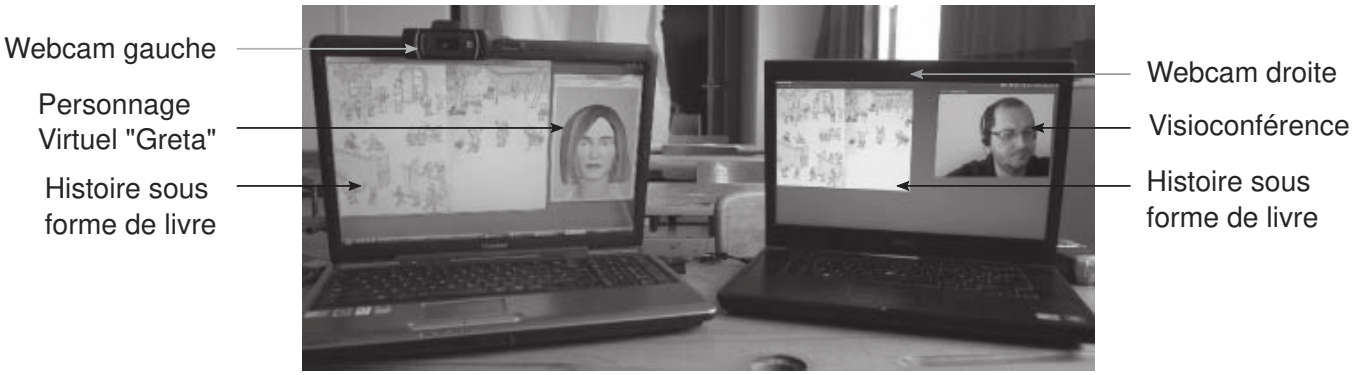

Figure 8. Vue du côté de l'enfant.

7. Résultats

L'étude présentée ici porte sur un ensemble de 20 enfants (7 filles et 13 garçons) dont la moyenne d'âge est de 7,7 ans (min : 6,11 et max : 8,1 ans) issus de deux écoles de Haute-Normandie. L'expérimentation se déroule au sein d'une école. L'enfant est amené dans la pièce, par un expérimentateur-accompagnant. Pendant la narration l'accompagnant est caché derrière un paravent. Ceci permet d'éviter toute référenciation sociale de l'enfant, en particulier lorsqu'apparaissent des erreurs interactionnelles. A la fin de la narration, l'expérimentateur, réalise un entretien avec l'enfant afin d'obtenir son ressenti concernant les trois narrateurs. Pour chaque enfant, l'expérience dure approximativement 20 minutes.

Les enfants de l'étude sont nombreux à utiliser les nouvelles technologies par le biais des jeux vidéos, 90% des enfants sont familiers avec les personnages animés et 60% avec l'utilisation de la webcam. En outre, des histoires sont lues à la majorité des enfants (86%), sachant que les enfants qui ne bénéficient pas de "l'histoire du soir" sont, en général, les enfants de 8 - 9 ans qui commencent à bien lire eux-mêmes (70% des enfants à qui on ne raconte pas d'histoire).

7.1. Grille d'analyse

L'analyse des interactions entre enfants et narrateurs a été réalisée à partir des vidéos enregistrées lors de la passation, via les webcams des deux ordinateurs, mais aussi sur la base des retranscriptions intégrales du discours des enfants. Les indices relevés concernent l'établissement d'une interaction par l'enfant avec le narrateur.

Dans un premier temps, nous nous sommes concentrés sur des indices sémantiques caractérisant les tours de paroles :

– Le nombre de mots, de phrases et le nombre de mots par phrase ont été comptabilisés à partir des retranscriptions.

– Les diffluences désignent l'ensemble des irrégularités, hésitations et reprises qui ponctuent le discours, rompent sa fluence et le rendent parfois plus complexe à comprendre (Oviatt, 2000). Dans ce travail, ont été comptabilisés les pauses remplies (ex : "*euh...*"), les faux départs (ex : "*je ne veux pas acheter... euh, prendre un ticket*"), les

autocorrections (ex : "*139... 1339*") et les répétitions (ex : "*de la... de la*"). Différentes études montrent que la fluence du discours varie en fonction de l'interlocuteur et de la situation de communication. Par ailleurs, les diffluences constituent une des difficultés majeures de retranscription automatique et de reconnaissance de la parole humaine en informatique (Oviatt, 2000).

Dans un second temps, nous avons réalisé une caractérisation de l'interaction entre les enfants et les deux narrateurs grâce à une analyse multimodale des réactions des enfants suite aux erreurs interactionnelles introduites dans le scénario :

– Les Mimiques Emotionnelles (ME) sont prises en compte en réaction aux incongruités insérées dans le scénario, en particulier les rires et les sourires mais aussi les mimiques de surprises ou d'incompréhensions exprimées par les enfants.

– La présence ou l'absence de Réponses Verbales Spontanées (RVS) est relevée. Ce sont toutes les prises de parole de l'enfant destinées à rompre le silence et à susciter la reprise de l'interaction ou toutes les réponses verbales des enfants suite aux erreurs interactionnelles et qui n'ont pas été suscitées par une question du narrateur.

L'annotation des réactions de l'enfant (différents niveaux de sourire, surprise, etc.) a été effectuée sur l'intégralité du corpus par 2 psychologues différents.

7.2. Engagement de l'enfant dans l'interaction

En situation d'interaction stimulée (suite à une des erreurs interactionnelles du scénario), les résultats issus de la comparaison du nombre moyen de réactions des enfants selon le narrateur montrent que les enfants communiquent davantage avec l'adulte filmé (m=2,2) qu'avec l'avatar (m=1,97), bien que l'effet de l'interlocuteur soit statiquement négligeable (h^2=0,02 ; ns, p=0,362) (voir figure 9).

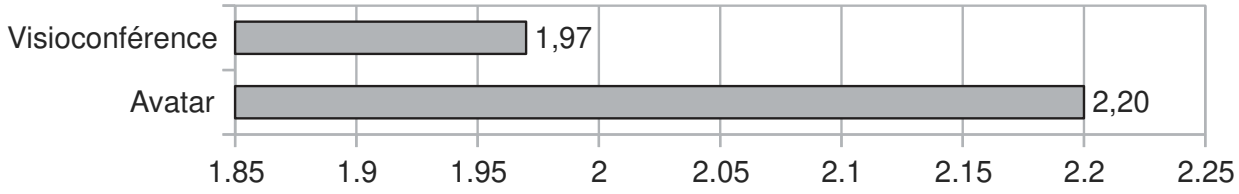

Figure 9. Moyenne des réactions (ME ou RVS) des enfants, en fonction du narrateur (personnage virtuel ou adulte filmé), suite à une erreur interactionnelle.

Ce résultat montre que les enfants considèrent le personnage virtuel comme un interlocuteur à part entière au même titre que l'humain. Ainsi, nous pouvons dire que les enfants acceptent la situation de narration interactive et ce quel que soit le narrateur. Les résultats suivants visent à préciser la nature de l'interaction selon le narrateur.

7.3. Nature de l'interaction

La figure 10 compare les réactions des enfants aux erreurs interactionnelles selon 4 familles de réponses : Les Mimiques Emotionnelles (ME), les Réponses Verbales Spontanées (RVS), les Réponses Verbales attendues après Question (RVQ) et les temps de Fixation Visuelle (FV). Cette évaluation montre que les enfants s'adressent différemment selon que le narrateur est un adulte filmé ou un avatar. En effet, les enfants communiquent préférentiellement avec l'avatar par des réponses verbales plutôt que par mimiques émotionnelles (s, p=0,014) alors qu'ils s'adressent préférentiellement à l'adulte filmé par le biais des mimiques émotionnelles (s, p=0,009).

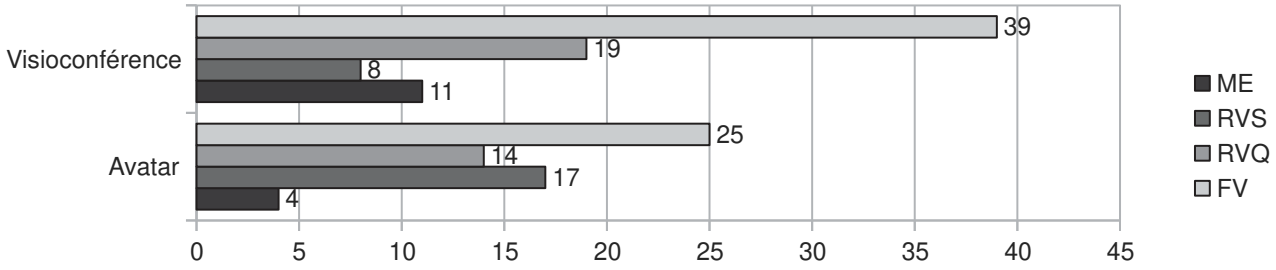

Figure 10. Nombre total de réactions des enfants aux erreurs interactionnelles, groupées par modalité de réponse.

Afin d'évaluer si cette prédominance de la modalité parlée pour l'avatar par rapport à l'adulte filmé existe pour l'intégralité des interactions et non uniquement pour les interactions suscitées, un ensemble de tests a été effectué. La figure 11 fournit des mesures quantitatives sur le nombre de phrases, de mots, de phrases interactives et de pauses (reprise de la parole de l'enfant après plus de 2 secondes). Ces mesures sont additionnées pour l'ensemble des 20 enfants. Les pauses, les phrases et les mots sont ceux prononcés spontanément par l'enfant, tandis que les phrases interactives sont déclenchées par le narrateur (adulte filmé ou avatar). Les données sont représentées pour les deux narrateurs : personnage animé et adulte filmé.

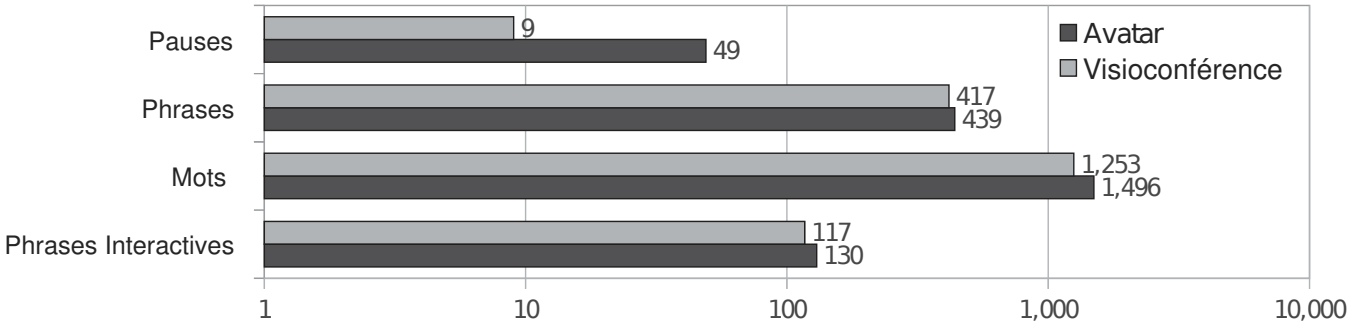

Figure 11. Nombre total de réactions des enfants aux erreurs interactionnelles, groupées par type de réponse.

Excepté pour le nombre de pauses, il n'existe pas de différence statistiquement significative pour la variable narrateur. Cela peut signifier que les enfants ne se sont pas saisis d'une différence entre les narrateurs. La seule différence significative est le nombre de pauses longues entre les questions des narrateurs et les réponses des

enfants. Nous avons vérifié qu'il ne s'agissait pas d'un déficit d'attention de la part de l'enfant, grâce notamment à l'enquête de fin de passation. Tous les enfants qui ont participé à l'expérience ont répondu correctement à l'enquête consistant à décrire certains éléments de l'histoire. Cette différence pourrait être expliquée par le style du narrateur. En effet, l'avatar a tendance à être plus monotone dans ces expressions (verbales et non verbales) que l'adulte filmé.

Enfin, la figure 12 présente les résultats issus de la comparaison de la longueur moyenne des phrases prononcées par les enfants avec chaque narrateur.

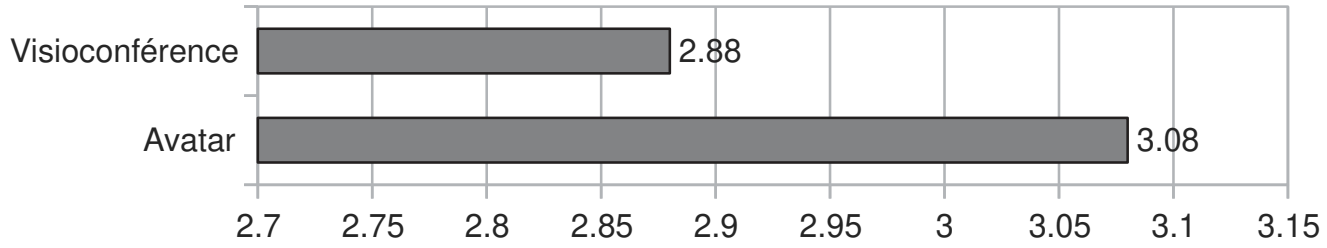

Figure 12. Longueur moyenne des phrases prononcés par les enfants par narrateur.

Ces résultats montrent que les enfants communiquent en utilisant des phrases plus longues avec l'avatar (m=3,08) qu'avec l'adulte filmé (m=2,88). Cependant l'effet de l'interlocuteur est statiquement négligeable (t=0,445; p=0,881>0,5; ns).

7.4. Réactions des enfants aux erreurs interactionnelles

En introduisant des erreurs interactionnelles dans le scénario, nous cherchons à provoquer l'interaction entre les enfants et les deux narrateurs. L'objectif est de sortir d'un contexte de narration pour entrer dans une situation de communication naturelle et spontanée. Trois types d'erreurs interactionnelles sont utilisées : les erreurs de Compréhension (C1 et C2), les Erreurs Emotionnelles (E1 et E2) et les erreurs d'Attention Conjointe (AC1 et AC2). La figure 13 compare ces types d'erreurs.

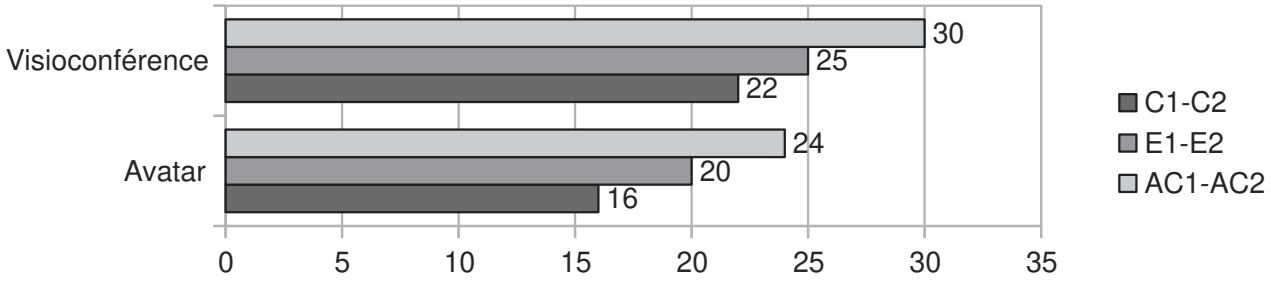

Figure 13. Le taux des réactions des enfants aux erreurs interactionnelles, par type d'erreur.

Dans le cas de l'erreur de compréhension C1, les enfants ont bien perçu l'erreur mais ne corrigent le terme "carotte" que si on leur pose directement la question (RVQ). Ils ne corrigent quasiment pas spontanément l'erreur (RVS). Par contre, dans le cas de

l'erreur C2, les enfants répondent à la question qui leur est posée mais pas de la façon attendue. C'est à dire qu'ils n'évoquent jamais le bruit qui couvre la voix du narrateur ou le fait qu'il n'ont pas compris le narrateur. Dans le cas de l'erreur C2, on note que les enfants, dans 55% des cas, se réfèrent au narrateur pour comprendre ce qui se passe en focalisant leur regard sur lui (ME, FV). Ils utilisent donc l'information visuelle.

Les entretiens de fin de passation ont mis en évidence que les erreurs AC1 et AC2 sont celles dont les enfants se souviennent plus spontanément. Le souvenir de l'erreur de compréhension C1 revient facilement quand elle est évoquée, ce qui est beaucoup plus rare pour C2. Les erreurs émotionnelles sont très rarement remémorées.

7.5. Diffluences

La figure 14 représente le taux de diffluences moyen pour 100 mots des enfants selon le narrateur.

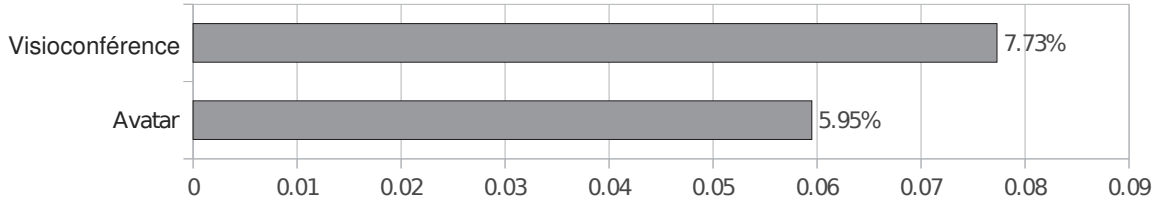

Figure 14. Taux de diffluences moyens pour 100 mots.

En comparaison des résultats obtenus dans l'étude d'Oviatt (Oviatt, 2000), le rapport entre les deux modalités (avatar-adulte filmé) est ici faible : le rapport est de 1,29 fois plus de diffluences avec l'adulte au lieu de 2,5 à 3 fois plus de diffluences observées dans l'étude d'Oviatt. Plusieurs hypothèses sont envisageables. Tout d'abord, lorsque l'avatar prend une forme humaine, le ratio de diffluences augmente. Deuxièmement, en raison du mode de vidéo-conférence, le ratio de diffluences est inférieur à ce qu'il serait en co-présence. En effet, dans son étude sur la fluence des dialogues entre adultes Oviatt mettait en évidence une augmentation du niveau de diffluences quand l'interaction avait lieu par téléphone (5,50 à 8,83). Il est donc possible que l'enfant, tout comme l'adulte dans l'étude d'Oviatt, modifie son comportement dans une communication médiée via une webcam.

7.6. Ressentis des enfants

Lors de l'entretien réalisé après l'expérimentation, les enfants ont eu des commentaires intéressants qui nous éclairent sur leur perception du personnage animé. Tout d'abord, les enfants ont repéré des différences entre les deux narrateurs, comme par exemple l'absence de microphone et de casque sur Poppy. D'autres ont comparé Poppy à un "*jouet*" ou à "*une dame en pâte à modeler*". En particulier, 55% des enfants ont remarqué une lenteur dans l'interaction avec l'avatar, mais aucun n'a pour autant trouvé cela gênant. Ainsi, tous les enfants se sont très bien adaptés au rythme de l'avatar et ont compris que cela leur permettait de parler. Cette adaptation facile

semble un gage de pouvoir utiliser un ACA, aux réactions plus lentes que celles d'un humain, sans que cela n'entrave trop l'interaction.

7.7. Synthèse des résultats obtenus

Nous pouvons conclure que les enfants sont capables de s'adapter au personnage virtuel et de profiter de l'interaction : le nombre d'interventions de l'enfant avec l'avatar ne diffère pas de celui avec l'adulte.

Il apparaît cependant que la communication avec l'avatar est de nature plus verbale (RVS) que non verbale (ME, FV) et ne semble pas aussi naturelle qu'avec un véritable narrateur (taux de diffluences moins élevé). Néanmoins, les enfants apprécient d'interagir avec l'avatar, comme l'a mis en évidence les entretiens de fin de passation. Il est intéressant de noter un discours moins hésitant, plus clair et plus assuré lorsque les enfants s'adressent au personnage animé. Ce sont des résultats qui intéressent les psychologues, les particularités du narrateur avatar pourraient être un avantage d'un point de vue thérapeutique, pour les enfants ayant des difficultés de communication.

Pour finir, en raison de l'interactivité de l'avatar, très peu d'enfants ont comparé Poppy avec un personnage de dessin animé. Ils interagissaient avec elle de façon similaire à ce qu'ils feraient avec un être humain, ce qui offre la possibilité de tester de nouveaux modèles de conversation sur les enfants. Le fait que la modalité verbale et sémantique soit préférée, même si non exclusive, avec l'avatar comme interlocuteur, laisse à penser que cette modalité porte plus d'importance dans le contexte d'un système d'interaction multimodale. Néanmoins, la transcription reste l'un des plus grands problèmes dans le domaine des systèmes interactifs.

8. Conclusion et perspectives

Nous avons montré dans cet article une méthodologie et des outils permettant d'améliorer la modélisation du dialogue. La méthodologie proposée consiste, d'une part, à extraire des motifs dialogiques afin de représenter les conventions dialogiques et, d'autre part, à effectuer une prédiction d'événements dialogiques afin de guider l'interaction avec l'auditeur. Une représentation matricielle de l'interaction est utilisée afin d'encoder les aspects multidimensionnels du dialogue.

Nos algorithmes ont été appliqués à un corpus de dialogues de narration parents-enfants afin d'en extraire un modèle de narration interactive. De plus, il ressort que la narration nécessite de nombreuses interactions et émotions.

Les interactions enfant-agent peuvent différer des interactions enfant-adulte, non seulement par les capacités dialogiques de l'agent, mais aussi plus simplement en raison de la représentation mentale que se font les enfants de l'agent. Afin d'évaluer l'impact de l'incarnation, une seconde expérience a été menée. Le modèle de dialogue narratif extrait précédemment a été utilisé durant une expérience de type Magicien d'Oz. Des dialogues de narration d'histoires enfantines enfant-adulte en visio-conférence

enfant-personnage virtuel (avatar piloté par un adulte) ont été collectés. Les résultats obtenus montrent que l'enfant est bien engagé dans l'interaction avec un agent virtuel mais que cet interaction se fait, avec le système présenté, davantage en utilisant la modalité orale que les gestes et mimiques faciales.

Nos travaux futurs se concentreront principalement sur l'intégration effective des modèles de dialogue obtenus au sein d'un ACA narrateur et à son évaluation en situation de narration interactive.

Remerciements

Ce travail a bénéficié du soutien du projet CNRS PEPS INS2I-INSHS « ACAMODIA ».

Bibliographie

Allen J., Perrault C. (1980). Analyzing intention in utterances. *Artificial Intelligence*, vol. 15, n° 3, p. 143-178.

André E., Pelachaud C. (2010). Interacting with embodied conversational agents. *Speech technology*, p. 123–149.

Antunes C. M., Oliveira A. L. (2001). Temporal data mining: An overview. In *Kdd 2001 workshop on temporal data mining.*

Astington J. W., Baird J. (2005). *Why language matters for theory of mind*. New York, Oxford University Press.

Aubergé V., Audibert N., Rilliard A. (2003). Why and how to control the authentic emotional speech corpora. In *Proc. of 8th european conference on speech communication and technology (eurospeech), geneva, switzerland*, p. 185–188.

Aust H., Oerder M., Seide F., Steinbiss V. (1995). The philips automatic train timetable information system. *Speech Communication*, vol. 17, n° 3-4, p. 249–262.

Bälter O., Engwall O., Öster A.-M., Kjellström H. (2005). Wizard-of-oz test of artur: a computer-based speech training system with articulation correction. In *Proceedings of the 7th international acm sigaccess conference on computers and accessibility*, p. 36–43.

Beale R., Creed C. (2009). Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies*, vol. 67, n° 9, p. 755–776.

Buisine S., Martin J.-C. (2003). Design principles for cooperation between modalities in bidirectional multimodal interfaces. In *Proceedings of the chi 2003 workshop on principles for multimodal user interface design, ft. lauderdale, florida.*

Bunt H. (2011). Multifunctionality in dialogue. *Computer Speech and Language*, vol. 25, n° 2, p. 222–245.

Cassell J. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents*, p. 1–27.

Cassell J., Bickmore T., Campbell L., Vilhjálmsson H., Yan H. (2000). Embodied conversational agents. In, p. 29–63. MIT Press.

Cavazza M., Camara R. S. de la, Turunen M. (2010). How was your day?: a companion eca. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems: volume 1-volume 1*, p. 1629–1630.

Chanoni E. (2009). Comment les mères racontent une histoire de fausses croyances à leur enfant de 3 à 5 ans ? *Enfance*, n° 2, p. 181-189.

Courgeon M., Clavel C., Martin J.-C. (2009). Appraising emotional events during a real-time interactive game. In *Affine'09*, p. 7:1–7:5. New York, NY, USA, ACM. http://doi.acm.org/10.1145/1655260.1655267

Douglas-Cowie E., Cowie R., Cox C., Amier N., Heylen D. (2008). The sensitive artificial listner: an induction technique for generating emotionally coloured conversation.

Dunn J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, vol. 3, n° 3, p. 32-57. http://www.tandfonline.com/doi/abs/10.1080/01969727308546046

Florek K., Lukaszewicz J., Perkal J., Steinhaus H., Zubrzycki S. (1951). Sur la liaison et la division des points d'un ensemble fini. In *Colloquium mathematicum*, vol. 2, p. 282–285.

Frampton M., Lemon O. (2009). Recent research advances in reinforcement learning in spoken dialogue systems. *Knowledge Engineering Review*, vol. 24, n° 04, p. 375–408.

Frey B., Dueck D. (2007). Clustering by passing messages between data points. *Science*, vol. 315, n° 5814, p. 972.

Gelin R., d'Alessandro C., Anh Le Q., Deroo O., Doukhan D., Martin J.-C. *et al. (2010). Towards a storytelling humanoid robot. In* Aaai fall symposium series.

Gouaillier D., Hugel V., Blazevic P., Kilner C., Monceaux J., Lafourcade P. et al. (2009). Mechatronic design of nao humanoid. *Proc. of the Int. Conf. on Robotics and Automation*, p. 769–774.

Grynszpan O., Martin J.-C., Nadel J. (2008). Multimedia interfaces for users with high functioning autism: An empirical investigation. *International Journal of Human-Computer Studies*, vol. 66, n° 8, p. 628–639.

Guha S., Rastogi R., Shim K. (2000). Rock: A robust clustering algorithm for categorical attributes. *Information Systems*, vol. 25, n° 5, p. 345–366.

Han J., Jo M., Park S., Kim S. (2005). The educational use of home robots for children. In *Robot and human interactive communication*, p. 378–383.

Hartholt A., Traum D., Marsella S. C., Shapiro A., Stratou G., Leuski A. *et al. (2013, août). All together now: Introducing the virtual human toolkit. In* International conference on intelligent virtual humans. *Edinburgh, UK.*

Hulstijn J. (2000). Dialogue games are recipes for joint action. In Proc. of gotalog'00.

Karypis G., Han E., Kumar V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer*, vol. 32, n° 8, p. 68–75.*

Klein J., Moon Y., Picard R. W. (2002). This computer responds to user frustration: Theory, design, and results. Interacting with computers*, vol. 14, n° 2, p. 119–140.*

Klemmer S. R., Sinha A. K., Chen J., Landay J. A., Aboobaker N., Wang A. (2000). Suede: a wizard of oz prototyping tool for speech user interfaces. In Proceedings of the 13th annual acm symposium on user interface software and technology*, p. 1–10.*

Kopp S., Krenn B., Marsella S., Marshall A., Pelachaud C., Pirker H. et al. (2006). Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent virtual agents*, p. 205–217.

Kozima H., Nakagawa C., Yasuda Y. (2005). Interactive robots for communication-care: a case-study in autism therapy. In *Robot and human interactive communication*, p. 341-346.

Lecroq T., Pauchet A., Chanoni E., Solano G. A. (2012). Pattern discovery in annotated dialogues using dynamic programming. *Int. J. of Intelligent Information and Database Systems*, vol. 6, n° 6, p. 603-618.

Lee J.-G., Han J., Whang K.-Y. (2007). Trajectory clustering: a partition-and-group framework. In *Int. conf. on management of data*, p. 593–604. ACM.

McKeown G., Valstar M. F., Cowie R., Pantic M. (2010). The semaine corpus of emotionally coloured character interactions. In *Multimedia and expo (icme), 2010 ieee international conference on*, p. 1079–1084.

McTear M. (2004). *Spoken dialogue technology: toward the conversational user interface*. Springer-Verlag New York Inc.

Miksatko J., Kipp K. H., Kipp M. (2010). The persona zero-effect: Evaluating virtual character benefits on a learning task with repeated interactions. In *Intelligent virtual agents*, p. 475–481.

Mori M. (1970). The uncanny valley. *Energy*, vol. 7, n° 4, p. 33–35.

Moundridou M., Virvou M. (2002). Evaluating the persona effect of an interface agent in a tutoring system. *Journal of computer assisted learning*, vol. 18, n° 3, p. 253–261.

Munteanu C., Boldea M. (2000). Mdwoz: A wizard of oz environment for dialog systems development. In *Lrec'00*.

Ochs M., Pelachaud C., Sadek D. (2008). An empathic virtual dialog agent to improve human-machine interaction. In *Proceedings of the 7th international joint conference on autonomous agents and multiagent systems-volume 1*, p. 89–96.

Otto M., Friesen R., Rösner D. (2011). Message oriented middleware for flexible wizard of oz experiments in hci. In *Human-computer interaction. design and development approaches*, p. 121–130. Springer.

Oviatt S. (2000). Talking to thimble jellies: Children's conversational speech with animated characters. In *Proc. icslp (beijing, china)*, p. 67–70.

Pelachaud C. (2009). Modelling multimodal expression of emotion in a virtual agent. *Philosophical Trans. of the Royal Society B: Biological Sciences*, vol. 364, n° 1535.

Poggi I., Pelachaud C., Rosis F., Carofiglio V., Carolis B. (2005). Greta. a believable embodied conversational agent. *Multimodal intelligent information presentation*, p. 3–25.

Prendinger H., Ishizuka M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, vol. 19, n° 3-4, p. 267–285.

Prendinger H., Mayer S., Mori J., Ishizuka M. (2003). Persona effect revisited. In *Intelligent virtual agents*, p. 283–291. Berlin Heidelberg.

Prepin K., Pelachaud C. (2013). Basics of intersubjectivity dynamics: Model of synchrony emergence when dialogue partners understand each other. In *Agents and artificial intelligence*, p. 302–318. Springer.

Robins B., Dickerson P., Dautenhahn K. (2005). Robots as embodied beings - interactionally sensitive body movements in interactions among autistic children and a robot. In *Robot and human interactive communication*, p. 54-59.

Ryokai K., Vaucelle C., Cassell J. (2003). Virtual peers as partners in storytelling and literacy learning. *Journal of computer assisted learning*, vol. 19, n° 2, p. 195–208.

Salber D., Coutaz J. (1993). Applying the wizard of oz technique to the study of multimodal systems. *Human-Computer Interaction*, p. 219–230.

Scaife M., Bruner J. S. (1975). The capacity for joint visual attention in the infant. *Nature*.

Schröder M. (2010). The SEMAINE API: towards a standards-based framework for building emotion-oriented systems. *Advances in HCI*, vol. 2010, p. 2–2.

Searle J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University.

Seif El-Nasr M., Wei H. (2008). Exploring non-verbal behavior models for believable characters. In *Interactive storytelling*, vol. 5334, p. 71-82.

Swartout W. R., Gratch J., Jr. R. W. H., Hovy E. H., Marsella S., Rickel J. *et al. (2006). Toward virtual humans.* AI Magazine*, vol. 27, n° 2, p. 96-108.*

Taymans W., Baker S., Wingo A., Bultje R., Kost S. (2001). Gstreamer application development manual.

Von Hofsten C., Rosander K. (2007). From action to cognition *(vol. 164). Elsevier Science.*

Von Luxburg U. (2007). A tutorial on spectral clustering. Statistics and Computing*, vol. 17, n° 4, p. 395–416.*

Whittaker S., Walker M., Moore J. (2002). Fish or fowl: A wizard of oz evaluation of dialogue strategies in the restaurant domain. In Language resources and evaluation conference.