Identification de Phénomènes dans l'analyse d'interactions humaines

Les traces d'interactions humaines, un nouveau domaine d'application pour la RI

Gregory Dyke[*], Michel Beigbeder[*], Kristine Lund[**], Jean-Jacques Girardot[*]

[*] *Ecole Nationale Supérieure des Mines de Saint-Etienne,*
158 Cours Fauriel, 42023 Saint-Etienne
[**] *UMR ICAR, CNRS, ENS-LSH, Université de Lyon,*
15 parvis René Descartes, F-69700 Lyon, France
{Gregory.Dyke, Michel.Beigbeder, Jean-Jacques.Girardot}@emse.fr
Kristine.Lund@univ-lyon2.fr

RÉSUMÉ. *L'étude socio-cognitive des interactions humaines médiatisées par ordinateur passe par l'analyse de corpus complexes, de plus en plus vastes, regroupant les enregistrements audio-video et les traces informatiques de l'interaction médiatisée. Dans cet article, nous présentons et modélisons l'interrogation de tels corpus au moyen de méthodes de RI. Nous montrons que, moyennant ces modèles, certaines questions de recherche pour l'analyse d'interactions peuvent se ramener à des problèmes connus de RI. Nous exposons enfin les résultats de nos premières implémentations d'algorithmes de RI pour l'interrogation de traces d'interaction.*

ABSTRACT. *The socio-cognitve study of human computer-mediated interactions can be done through the analysis of increasingly larger and complex corpora composed of audio-video recording and interaction logfiles. In this article, we present and model the querying of such corpora with IR methods. We show that these models afford the transformation of certain interaction analysis research questions into known IR problems. We describe the results of our first implementations of RI algorithms for querying interaction corpora.*

MOTS-CLÉS : *Analyse de traces, Traces d'interaction, Système dédié*

KEYWORDS: *Interactive IR and visualization, Non-probabilistic retrieval models, Domain-specific search engines, Interaction traces*

CORIA, pages 1 à 12

1. Introduction

Les études socio-cognitives des interactions humaines supportées par ordinateurs s'effectuent aujourd'hui le plus souvent au travers des traces de ces activités (*logs* des outils, enregistrements numériques audio ou vidéo). L'analyse de ces traces est complexe, et les données sont en trop grande quantité pour être facilement appréhendées (Dyke *et al.*, 2007), même si depuis peu quelques techniques issues de l'état de l'art informatique (p.ex. apprentissage automatique pour assister le chercheur dans le codage de ses données (Erkens *et al.*, 2008)) ont été mises en oeuvre pour faciliter la tâche d'analyse. L'utilisation de l'informatique reste cependant limitée malgré les résultats prometteurs de chercheurs utilisant cet outil pour diverses tâches (Cox, 2007).

Le temps passé par un chercheur à analyser une séquence de données est de l'ordre de 10 et 1000 fois la longueur de cette séquence [1]. Il est dès lors d'une importance capitale pour lui de pouvoir repérer rapidement dans un grand *corpus* [2] les *phénomènes*[3] qui l'intéressent. Nous nous sommes donc proposés d'adapter et d'appliquer certaines approches de la RI à l'analyse de corpus, pour permettre au chercheur de décrire, sous forme de requêtes, les aspects des phénomènes qu'il recherche, et retrouver ainsi, de manière automatique, ces phénomènes potentiels.

Dans cet article, nous passons en revue l'activité d'analyse de ces chercheurs afin de souligner les apports potentiels de la RI, et présenter les modèles de représentation de corpus et d'interrogation qui permettent de tirer parti des techniques évoquées. Nous décrivons enfin *Tatiana*, un outil destiné à aider l'analyse de traces, qui implémente certaines des fonctionnalités décrites, et montrons quelques résultats obtenus sur des corpus réels.

2. L'analyse d'interactions humaines médiatisées par ordinateur

Les raisons d'être des analyses de corpus d'interactions médiatisées par ordinateur sont nombreuses. Le but peut être la mise au point d'un outil de CSCL/CSCW (Computer Supported Cooperative Learning/Work), l'analyse de son ergonomie ou l'évaluation de ses procédures d'usage. Dans le cadre du CSCL, de très nombreux logiciels sont utilisés, des plus généraux (gestionnaires de chats, de forums) aux plus spécialisés (micro-mondes, outils de conception collaboratifs). L'évaluation d'un apprentissage effectué par des étudiants au cours d'une session s'effectue souvent au travers du calcul d'indicateurs. L'enseignant s'intéressera par exemple aux connaissances manipulées, le didacticien au transfert de ces connaissances, le cogniticien à

1. Chiffres tirés de discussions avec des chercheurs en science cognitives, en particulier lors des séminaires internes de l'UMR ICAR.
2. Nous utilisons tout au long de cet article le terme *corpus* dans l'acceptation très particulière que lui prêtent les socio-cogniticiens, pour désigner l'ensemble des documents relatifs à une observation. La section 3 en présente une description précise.
3. Au sens de la psychologie cognitive, c'est à dire un changement dans l'enchaînement habituel d'un processus social ou cognitif.

l'argumentation ou à la reformulation qui sont à l'origine de ce transfert, etc. Dans ces exemples, les captures audio et vidéo des expérimentations sont indispensables à la compréhension des interactions (Goodman *et al.,* 2006, Avouris *et al.,* 2007). Les outils de CSCW sont aussi de plus en plus utilisés dans l'industrie, entre autres comme support de réunions à distance. Les traces de ces outils constituent ainsi une mémoire de l'entreprise, dont l'importance est croissante, car de plus en plus de décisions administratives ou techniques sont prises dans de telles circonstances (Lund *et al.,* 2007).

Nous prenons essentiellement dans cet article des situations dans le domaine de l'analyse socio-cognitive des traces d'interactions, puisque la plupart des cas sur lesquels nous avons travaillé rentrent dans cette spécialité. Illustrons ce domaine au travers d'un premier exemple typique (*cas 1*) : un chercheur en cognitique s'intéresse à la manière dont la *reformulation*[4] participe au transfert des connaissances entre enseignant et étudiants dans le cadre d'un suivi de projets d'un cours d'informatique. Dans le cas étudié, trois rencontres, d'une durée d'une heure chacune, se déroulent à une semaine d'intervalle environ. Deux enseignants et dix binômes d'étudiants sont impliqués dans l'expérimentation, soit 30 heures de session et 22 participants différents. Les rencontres ont lieu en présentiel, les intervenants utilisant soit l'oral, soit un outil de communication offrant un forum, un éditeur de texte partagé, un tableau blanc, un dispositif de construction de graphes argumentatifs, etc.

A partir des données brutes (traces des outils et enregistrements audio et vidéo), le chercheur va préparer le corpus pour obtenir des *artefacts*[5] primaires présentant la granularité désirée. Ce travail implique de synchroniser les différents médias, de convertir les traces dans une représentation appropriée à l'outil d'analyse, et de regrouper ou fragmenter les événements primaires pour les amener à la granularité souhaitée. L'audio et la vidéo sont segmentés en tours de paroles, qui sont transcrits manuellement. Les événements des traces d'interactions sont regroupés, pour obtenir des interventions complètes dans le chat, des séquences signifiantes dans l'éditeur de texte, etc.

Dans l'approche traditionnelle du corpus, l'examen de celui-ci est entièrement « manuel » : le chercheur visionne de manière répétitive les séquences du corpus, pour tenter de découvrir les phénomènes qu'il recherche afin d'appuyer son analyse. Ainsi, le phénomène de « reformulation » va être caractérisé par le fait qu'un enseignant utilise oralement un terme ou un groupe de termes, que l'on retrouve peu après sous la forme d'une prise de notes, par un étudiant, dans le chat ou l'éditeur partagé. Muni de ces exemples de reformulation, le chercheur va ensuite tenter de montrer que les étudiants se sont appropriés les connaissances ainsi reformulées.

Dans une autre étude (*cas 2*), un chercheur nous avait soumis un ensemble de

4. Le terme de *reformulation* est utilisé ici non pas dans sa signification psychanalytique traditionnelle, mais pour désigner un phénomène dans lequel un locuteur B reprend, sous une forme voisine, sur le même support ou un support différent (oral-oral, oral-écrit, etc.), un concept qui vient d'être énoncé par un locuteur A.
5. Nous utilisons ici ce terme dans son acceptation anglo-saxonne, pour indiquer un objet artificiel créé par le chercheur, et destiné à l'aider dans son analyse.

requêtes, correspondant aux phénomènes qu'il recherchait dans plusieurs corpus de transcriptions orales, requêtes dont voici quelques exemples :

– chevauchement (deux locuteurs parlent simultanément) suivi de « attends » dans les PV (productions verbales) de moins de dix tokens,

– cascade de trois « attends » dans un intervalle de cinq PV maximum,

– fréquence de « oui » par sexe dans les enregistrements avec des locuteurs mixtes,

– PV avec chevauchement et « attends » précédées d'une PV avec « euh », etc.

Un autre exemple (*cas 3*) a été l'analyse d'une situation réelle du monde industriel, mettant en scène des concepteurs de métiers différents effectuant une réunion dont le but était de ratifier les décisions prises au sujet de 12 points de conception. Lors de l'analyse, les chercheurs désiraient identifier des phénomènes communs entre la façon dont est abordée chacun de ces points. A ces fins, ils ont transcrit le corpus et l'ont codé en fonction du type des énoncés (affirmations, questions, etc.) et de leur contenu (orienté-solution, orienté-technologie, etc.)

On voit à travers ces exemples que la détection des phénomènes intéressant le chercheur relève de la recherche de passages dans des documents structurés, avec utilisation d'un aspect de proximité temporelle. De manière générale, nous attendons de l'utilisation de la RI dans la détection de phénomènes qu'elle réduise le nombre de cycles nécessaires à cette détection, qu'elle en diminue le silence, qu'elle raccourcisse le temps consacré par le chercheur à cette recherche, et rende possible l'application d'une question de recherche sur des corpus de très grande taille.

3. Modèle de la représentation

Le terme de *corpus* désigne ici l'ensemble des documents que le chercheur veut étudier comme un tout. Dans l'exemple évoqué, il s'agit des documents recueillis au cours d'une expérience, qui consistent typiquement en :

– documents audio ou vidéo enregistrés pendant l'expérience, transcriptions de ces enregistrements effectuées par le chercheur (on peut considérer aussi que ces transcription constituent un artefact secondaire) ;

– traces des interactions médiatisées ;

– notes prises avant, pendant, et après l'expérience par les participants ou les observateurs ; documents distribués aux participants de l'expérience, et, de manière générale, tout autre document jugé pertinent par le chercheur.

Ces documents constituent un ensemble fini, le « corpus primaire », qui n'a pas de raison d'évoluer *a posteriori*, représentant la totalité des données primaires recueillies pendant et sur l'expérience. Cet ensemble de documents est assez disparate. Les outils de CSCL/CSCW sont nombreux, ont souvent des finalités différentes, et utilisent bien sûr des représentations différentes pour leurs traces. Certaines expérimentations peuvent en utiliser plusieurs simultanément. Nous ne pouvons envisager un outil gé-

nérique, capable d'analyser tous les formats disponibles, et ceux à venir, pas plus qu'il n'est raisonnable d'espérer proposer un jour un format « universel » pour ces données, susceptible de représenter toutes les informations disponibles dans ces différentes traces.

Nous prenons comme hypothèse que les données primaires sont immuables, mais que le chercheur va travailler sur des « représentants » de ces données primaires. L'approche choisie a été de créer un *format pivot*, pouvant représenter toutes les informations auxquelles le chercheur peut s'intéresser durant son analyse. Ce format permet au chercheur de choisir les aspects des données recueillies qu'il souhaite analyser, en laissant les autres éléments de côté. Ce choix n'est en rien irréversible : à tout moment, il est possible d'ajouter de nouveaux aspects, extraits du corpus primaire, sans perdre aucun résultat du travail effectué. Il est de même possible à deux chercheurs de travailler sur un même corpus, en s'intéressant à des données qui ne se recouvrent pas totalement, puis ultérieurement de croiser leurs analyses, pour faire apparaître de nouveaux phénomènes.

La représentation choisie pour le format pivot est une représentation XML qui utilise une structure de donnée, que nous avons baptisée *item*, et qui n'est pas sans évoquer les *frames* (Minsky, 1974). Un item est une représentation d'un objet du monde réel modélisé sous la forme d'un ensemble de facettes, chaque facette étant un couple nom-valeur décrivant un aspect de l'objet. Chaque item créé comme représentant d'un objet du corpus comporte une facette qui décrit sa provenance, facette dite *ancre*. Pour un document XML présent dans le corpus, par exemple, l'ancre est constituée de l'identification du document et du chemin d'accès à l'élément d'où l'on a extrait l'information.

Le type d'item le plus important utilisé dans l'analyse est l'*événement*. Un tel item comporte une *datation* de l'événement, sous la forme d'un couple date-durée. Les autres facettes de l'item documentent d'autres aspects de l'événement auquel s'intéresse le chercheur, tels que l'outil utilisé, l'identification du participant, la description de l'action, etc. Un message envoyé dans un chat pourra typiquement être représenté par un événement contenant des facettes "outil", "participant" et "message", en plus des facettes "ancre" et "datation". La trace primaire de l'outil peut contenir nombre d'autres informations (le numéro du message, du groupe, l'adresse IP de la machine, etc.), que le chercheur peut juger, ou non, pertinentes, et associer ou non à l'événement. La trace de l'utilisation d'un outil est ainsi traduite par une successions d'événements, qui est représentée, dans notre formalisme, par une séquence d'items. Une telle séquence chronologique d'événements associée à un outil est dite *rejouable*.

La transformation de la trace d'un outil quelconque de CSCL/CSCW en une séquence d'items s'effectue en général par un script *ad hoc* écrit en XQuery (W3C 2008) pour des traces XML, ou en Java pour les autres représentations. Cette approche permet au chercheur de ne conserver pour son analyse que les aspects des données disponibles qui l'intéressent, tout en affranchissant l'outil d'analyse de la prise en compte de formalismes différents. Pour un logiciel particulier, une dizaine de scripts de quelques lignes suffisent en général à couvrir tous les besoins du chercheur.

4. Modèle de l'analyse

Les événements recueillis au cours de l'expérimentation ont rarement la granularité adéquate à l'analyse. Le chercheur va éliminer des événements qu'il juge non significatifs, ou regrouper plusieurs événements successifs, afin d'obtenir une « action » de plus haut niveau sémantique. Ainsi, il peut choisir de supprimer une succession de déplacements d'un objet dans un tableau blanc, grouper une série d'insertion de caractères dans un éditeur partagé pour obtenir un événement qu'il qualifiera d'insertion de mot ou de phrase, etc. Il va aussi catégoriser ces événements, leur associer des commentaires, ou les lier à d'autres objets du corpus. Toutes ces opérations se traduisent par la création de nouveaux artefacts, qui concrétisent des vues spécifiques sur le corpus, ou certains aspects de la réflexion du chercheur.

Dans cette étape du travail, les affichages de données prennent toute leur importance. Le chercheur peut utiliser, en succession ou simultanément, des affichages sous forme de tables (à la manière d'Excel), de graphes, de représentations linéaires temporelles, etc. Il peut décider d'effectuer de nouvelles transformations sur ces données, leur ajouter de nouvelles facettes extraites du corpus primaire ou d'artefacts secondaires, les enrichir ou les filtrer de diverses manières. Un atout considérable pour le chercheur consiste aussi à pouvoir rejouer, avec l'outil ayant servi à produire les traces qu'il analyse, les traces elles-mêmes, et voir se dérouler sous ses yeux, en temps réel, ce que pouvaient voir sur leurs écrans les participants à l'expérience. Pour cette raison, le chercheur privilégie les outils permettant de rejouer les sessions enregistrées, tels que DREW (Corbel *et al.,* 2002), Digalo (Lotan-Kochan, 2006), CoFFEE (De Chiara *et al.,* 2007) et quelques autres.

Le travail du chercheur est donc caractérisé par une succession de cycles, au cours desquels il analyse (manuellement) ses données ou leurs représentations, crée de nouveaux objets de travail, puis en effectue à nouveau l'étude. La phase durant laquelle le chercheur examine ses rejouables pour trouver des évidences des phénomènes qu'il cherche à mettre en lumière, est, on le voit, la partie critique, en temps et en effort humain, du processus d'analyse.

5. Quels outils de la RI pour l'analyse de traces ?

L'identification des phénomènes recherchés par le chercheur, et la création interactive de nouveaux rejouables qui représentent ces phénomènes, est une des opérations les plus gourmandes en temps. C'est lors de cette phase d'examen des documents que diverses approches proposées par la RI peuvent s'avérer pertinentes.

La recherche d'information, au sens de recherche de documents, est représentée du point de vue du système comme une activité en deux phases : la phase d'indexation et la phase d'interrogation. La phase d'indexation explicite les entités qui seront manipulables au moment de l'interrogation. Dans un système classique de recherche d'information dans un corpus de documents textuels, le texte est découpé éléments (termes, mots, formes, *token* en anglais) dont on conserve les occurrences d'appari-

tion à un certain niveau de granularité (document, position dans le texte, position dans la structure du document) pour permettre un accès efficient au moment de l'interrogation.

Dans notre problématique, nous ne nous intéressons pour le moment qu'à des traces d'interactions en français ou en anglais, et pouvons donc nous appuyer sur la notion de *mot*. (Cependant, dans le cas d'un éditeur de texte partagé, un mot peut se retrouver divisé entre deux événements distincts parce que le système de centralisation met à jour ses données sur un niveau de granularité temporel qui n'est pas relié au niveau sémantique de la notion de mot ; pour que le mot soit retrouvé au moment de l'interrogation, il faut soit grouper les événements au moment de l'indexation en tenant compte des natures des médias, soit retarder ce travail jusqu'au moment de l'interrogation, ce qui est plus pertinent, dans la mesure où l'on travaille sur des artéfacts créés dynamiquement).

La deuxième notion manipulée par un système de recherche d'information est celle de *document*. C'est au moment de l'indexation que cette notion est explicitée au système. En gardant la définition de *document* comme unité retournée par le système, il faudrait donc que les documents soient les phénomènes. Le problème est que ce qui est (sera) un phénomène va parfois être découvert par le chercheur lui-même, au fur et à mesure de son interaction avec le système, et ne peut dès lors pas toujours être connu au moment de l'indexation.

Nos besoins en informations semblent plus proches de la recherche de passage, focalisée sur la recherche d'extraits de documents qui traitent du sujet évoqué par la requête. Les passages peuvent être définis de multiples façons : unités linguistiques, unités sémantiques (Hearst, 1997), unités lexicales, unités structurelles, etc. Les techniques employées consistent à définir une notion de passage, puis expliciter ceux-ci et enfin à appliquer des techniques de RI *ad hoc* où les documents sont les passages ainsi définis. Quelques travaux se sont focalisés sur définir un modèle de recherche *ad hoc* basé sur une recherche de passage (Wilkinson, 1994).

Dans notre cas, définir les passages n'est pas plus possible que de définir les documents lorsque les phénomènes ne sont connus qu'une fois le travail du chercheur terminé. Il faut donc se tourner vers une branche de la RI où ce qui est cherché est défini dans la requête elle-même. Les travaux qui se placent dans cette branche sont pour la plupart reliés à une notion de structure, en particulier les travaux autour des langages de requêtes dans les documents XML. Trois langages de requête correspondent actuellement à des standards actifs : NEXI, XPath et XQuery (W3C 2008). NEXI, initiative d'INEX (INEX, 2008), présente certaines extensions pour la recherche de texte au sein d'éléments XML, mais n'est pas assez puissant en termes de manipulation des autres types de données. XPath est d'abord un langage de désignation au sein d'un document XML, et constitue de fait un simple sous-ensemble de XQuery dans la version 1.1 de ce langage. XQuery propose la quasi-totalité des outils nécessaires, mais n'est pas complètement adapté à notre problème dans la mesure où il ne permet de retrouver que des éléments de la structure de base des documents, alors que les phénomènes ne sont pas prédéfinis et qu'il n'y a aucune raison pour qu'ils coïncident

avec un élément, d'autant que, comme nous l'avons indiqué, la structure que nous manipulons, bien qu'étant du XML, ne reflète pas la structuration sémantique de haut niveau.

Notre besoin est de pouvoir définir, dans des requêtes, des regroupements d'événements qui vérifient un certain nombre de contraintes. L'algèbre de recherche dans les textes structurés définie par (Clarke *et al.*, 1994) se rapproche de nos besoins. S'appliquant à du texte pur, cette algèbre définit la notion d'ensemble de recouvrement d'un ensemble de termes. Dans le cas de documents structurés, les balises apparaissent au même niveau que les mots, et sont traitées quasiment de la même manière, de l'indexation à l'interrogation. Leurs apparitions créent une partition exacte de l'ensemble du texte, définissent ainsi les documents. Avec des balises qui indiquent le locuteur et le texte qu'il dit (que nous appelons *PV*, *production verbale*), la combinaison texte-structure permet de repérer, par exemple, des intervalles contenant une *PV* dite par *A* et contenant *a,* suivie d'une *PV* dite par *B* et contenant *b*. Ceci nous rapproche des besoins de notre problème où le chercheur doit pouvoir poser des requêtes en désignant les intervenants et des termes utilisés dans leurs interventions. Cependant, une notion qui n'est pas prise en compte par le formalisme de l'algèbre de Clarke et al. est le parallélisme entre plusieurs documents. Dans nos corpus, le regroupement des événements et leur annotations amènent des séquences en parallèle de séquences déjà existantes (par exemple, transcription, chat, editeur de texte).

Enfin une dernière fonctionnalité dont ont besoin les chercheurs est la généralisation des requêtes. Si le formalisme précité permet d'interroger la base en instanciant toutes les données, il ne permet pas de créer des requêtes quantifiées, c'est-à-dire de poser des requêtes avec des variables : par exemple, chercher des *PV* de *X* contenant *x,* suivies d'une *PV* de *Y* (*Y* différent de *X*) et contenant *x*. Cette fois *X*, *Y*, et *x* ne sont plus des instances mais des variables et une réponse devra non seulement fournir des intervalles, mais aussi instancier ces variables. On voit donc que les besoins vont au delà de ce que traite la recherche d'information traditionnelle, même si les prémisses sont les mêmes.

6. Aspects de l'implémentation

Il n'est pas possible de détailler ici les différents aspects du modèle implémenté dans Tatiana. Un rapport de recherche, à paraître en 2009 devrait en donner les principaux aspects, ainsi qu'une sémantique opérationnelle. Le modèle proposé associe différentes propriétés aux rejouables, l'une d'entre elles étant la *traçabilité*, qui permet, à partir de n'importe quel élément d'un rejouable, de retrouver sa source dans le corpus initial. Le modèle décrit différentes transformations applicables à un rejouable, et propose en particulier trois opérations de recherche d'information au sein d'un rejouable S :

$l\ f \vdash s\ S$ (*R-1*) fournit la séquence des items de S dont les facettes, sélectionnées par s, contiennent tous les éléments de l ; le résultat est une séquence d'items I_i, ordonnés comme ceux de S, tels que $s(I_i)$ contient les éléments de l.

$l\ f \Vdash s\ S$ (*R-2*) fournit une séquence de résultats, dont chacun est une séquence R_i d'éléments I_j, tels que chaque $s(I_j)$ contienne au moins un élément de l, et que l'union des $s(I_j)$ contient tous les éléments de l.

$l\ f \Vdash\!\!\vdash s\ S$ (*R-3*) fournit une séquence de résultats, dont chacun est une séquence R_i d'éléments I_j, tels que l'union des $s(I_j)$ contient tous les éléments de l.

Informellement, l'objet l est une séquence (ordonnée) ou un sac (sans ordre) d'objets, typiquement des chaînes de caractères représentant des mots, f un prédicat de comparaison, et s une fonction de sélection d'une facette spécifique des items ; l'opération prend en charge l'itération sur cette facette des items, et recherche les objets de l, soit (*R-1*) dans un item unique (opération de recherche classique, la cible étant l'item), soit (*R-2*) dans une séquence d'items (chaque item contenant au moins l'un des objets), soit encore (*R-3*) dans un média continu obtenu par la fusion de l'ensemble des facettes. Dans tous les cas, le résultat est l'ensemble des séquences d'items contenant l'ensemble des objets recherchés.

De manière générale, le modèle proposé permet de traduire une requête complexe en un algorithme basé sur les opérations de transformations (sélections, tris, recherches, fusions, unions, etc.) appliquées à un rejouable. Cette requête peut s'effectuer de manière automatique, dispensant dans certains cas le chercheur de procéder à des regroupements manuels, et simplifiant la visualisation et la découverte des phénomènes qu'il recherche.

7. Résultats

Dans notre cas d'étude 3 (et suite à des tentatives infructueuses de fouille de données), nous avons souhaité donner le moyen aux chercheurs d'identifier des phénomènes constitués d'une séquence d'interactions, par exemple d'identifier les passages où tel locuteur répond à une question de tel autre locuteur. Cette recherche passe par une série de transformations, dont l'utilisation de la fonction *R-2* sur les facettes « locuteur » ou « texte ». Afin d'augmenter l'efficacité de ces opérations de recherche, nous effectuons une indexation partielle de séquences (indexer toutes les séquences sur un corpus n'est pas une opération pratiquable, puisqu'elle serait en $O(n^n)$; cependant, la pertinence d'une séquence étant liée à la distance entre son premier et dernier élément, il s'est avéré utile d'indexer toutes les séquences sur une fenêtre de 1 à *m* éléments, nos expérimentations nous ayant montré que, sur ce corpus particulier, une valeur de $m = 15$ donnait des résultats significatifs ; ceci revient, implicitement, à introduire une notion de proximité).

Ayant effectué cette indexation, il s'avère que la comparaison entre l'index simple des termes (implémentation de l'indexation pour *R-1*) et celle des séquences peut nous

donner des informations sur l'espérance d'occurence de ces séquences. En comparant cette espérance avec les occurences réelles, nous pouvons déterminer une *mesure de surprise* d'une séquence donnée S, que nous calculons par :

$$\frac{|E_S - O_S|}{E_S}$$

(où E_S est l'espérance d'occurences de S si les événements étaient distribués selon leur fréquence d'apparition dans l'ensemble étudié, et O_S le nombre effectif d'occurences observées de la séquence S) afin de trier les séquences identifiées par ordre de surprise. Nous avons ensuite évalué la pertinence de ces séquences « suprenantes » en les soumettant à l'approbation des chercheurs dont l'analyse de cette réunion a occupé une part importante des deux dernières années. Certaines de nos trouvailles étaient immédiatement évidentes : nous avons identifié une séquence (pour laquelle notre mesure donnait une valeur particulièrement importante), où l'un des locuteurs intervenait à plusieurs reprises, alors que ce locuteur n'avait eu que 5 tours de parole sur une réunion de 90 minutes et que ces tours avaient été presque consécutifs. D'autres de ces trouvailles étaient connues - mais seulement grâce au fait que les chercheurs avaient passé énormément de temps (entre 200 et 300 heures) à analyser ce corpus : le fait qu'un des locuteurs parlait souvent chaque fois que la discussion abordait un nouveau point - en effet il se chargeait alors de résumer les enjeux relatifs à cette décision. Enfin, nous avons identifié des séquences surprenantes que nos collègues n'avaient pas détectées, et qui leur ont apporté une vision nouvelle sur le corpus : ainsi, le fait que le locuteur *A* et le locuteur *B* n'intervenaient presque jamais après le locuteur *C* leur a permis de mettre en évidence une distinction entre deux sous-groupes de participants dans la réunion, et de quantifier cette distinction. Ce dernier résultat a été particulièrement utile pour mettre en évidence la structure interactionnelle complexe d'une réunion qui ne semblait pas présenter de structure spécifique à ce niveau. Notons encore que certaines séquences ainsi découvertes ont été effectivement reconnues comme étant des « phénomènes » par les chercheurs, même s'ils ne parviennent pas encore à en donner d'interpétation.

Dans le cas 1, évoqué dans la section 2, le chercheur souhaitait identifier la reformulation entre un dialogue oral et des notes écrites. Il s'agissait donc simplement d'utiliser la RI pour faire de la recherche de passages en utilisant les éléments rédigés comme requêtes, afin de trouver ces passages dans la transcription. L'un des avantages majeurs de l'application de la RI pour nos collègues a été la possibilité d'identifier *tous* les passages sources potentiels. En effet, au premier abord une phrase particulière dans les notes provient probablement d'un énoncé de la transcription relativement proche dans le temps. Nous avons cependant pu mettre en évidence que la formulation de la note pouvait parfois provenir de plusieurs énoncés, et que ce qui semblait être une reformulation relativement majeure d'un énoncé était en réalité la combinaison de cet énoncé avec quelque chose qui avait été dit un quart d'heure plus tôt (cf. fig. 7.1).

Chaque chercheur désirant identifier des phénomènes différents, qui ont des caractéristiques différentes pour la modélisation de la requête, nous ne sommes pas encore en mesure de présenter tous les avantages de l'application de la RI aux problèmes de

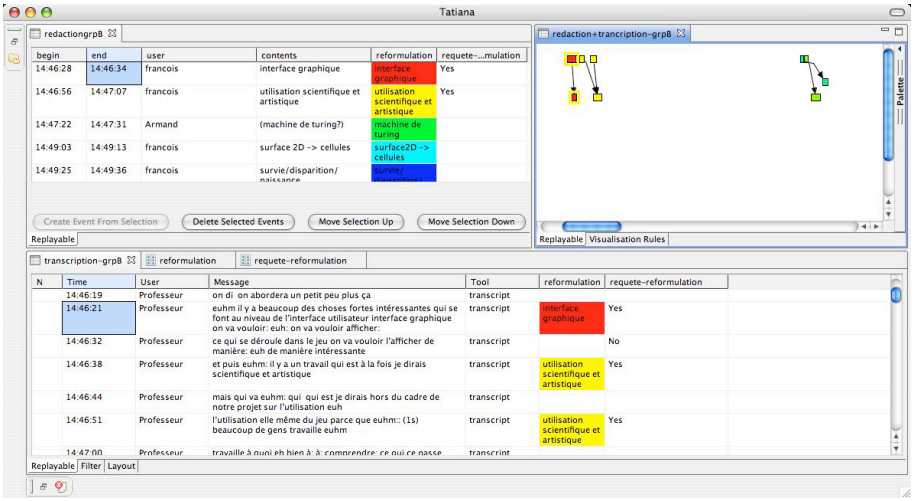

Figure 7.1. *Une capture d'écran du logiciel Tatiana. En haut, partie gauche, un affichage d'événements (après transformations) de l'éditeur de texte. En bas, la transcription après segmentation. En haut, partie droite, une visualisation temporelle linéaire des liens de reformulation trouvés par Tatiana.*

l'analyse des traces d'interaction. Nous nous sommes limités à implémenter quelques algorithmes classiques sur notre modèle de corpus, associés à un ensemble d'opérations de transformation des données, afin que ces premiers outils soient à la disposition des chercheurs souhaitant analyser leurs données. Plus d'une douzaine d'analyses sont actuellement en cours et utilisent ces outils, montrant leur pertinence pour ce domaine d'application.

8. Conclusions

Dans cet article, nous avons présenté le travail d'analyse dans une perspective socio-cognitive des interactions humaines médiatisées par des outils de CSCL et de CSCW, en montrant certaines des difficultés auxquelles sont confrontés les chercheurs désirant analyser des corpus composés de traces de telles interactions. Nous avons évoqué un modèle pour la représentation des corpus et pour leur interrogation, qui permet de rapprocher les problèmes d'analyse des chercheurs de problèmes de RI déjà connus. Le CSCL/CSCW acquiert un outil performant permettant la maitrise et l'analyse de corpus toujours plus grands, permettant aux chercheurs de travailler plus vite, et même d'identifier des phénomènes qui n'auraient pas été découverts manuellement.

9. Bibliographie

Avouris N., Fiotakis G., Margaritis M., Komis V., « Beyond logging of fingertip actions : analysis of collaborative learning using multiple sources of data. », *Journal of Interactive Learning Research*, vol. 18, n° 2, p. 231-250, 2007.

Clarke C. L., Cormack G., Burkowski F., An Algebra for Structured Text Search and A Framework for its Implementation, Technical Report n° CS-94-30, Dept. of Computer Science, Waterloo, Canada, August, 1994.

Corbel A., Girardot J.-J., Jaillon P., « DREW : "A Dialogical Reasoning Web Tool" », *ICTE2002, Intl. Conf. on ICT's in Education*, Badajos, Espagne, November, 2002.

Cox R., « Technology-enhanced research : educational ICT systems as research instruments », *Technology, Pedagogy and Education*, vol. 16, n° 3, p. 337-356, 2007.

De Chiara R., Di Matteo A., Manno I., Scarano V., « CoFFEE : Cooperative Face2Face Educational Environment », *Proceedings of the 3rd International Conference on Collaborative Computing : Networking, Applications and Worksharing*, New-York, USA, 2007.

Dyke G., Girardot J.-J., Lund K., Corbel A., « Analysing Face to Face Computer-mediated Interactions », *Developing Potentials for Learning*, Budapest, Hungary, august, 2007.

Erkens G., Janssen J., « Automatic coding of communication in collaboration protocols », *International Journal of Computer-Supported Collaborative Learning*, vol. 3, n° 4, p. ?, 2008.

Goodman B. A., Drury J., Gaimari R. D., Kurland L., Zarrella J., Applying User Models to Improve Team Decision Making, Technical Report n° 1351, Mitre, mitre.org, 2006.

Hearst M. A., « TextTiling : segmenting text into multi-paragraph subtopic passages », *Comput. Linguist.*, vol. 23, n° 1, p. 33-64, 1997.

INEX, « Initiative for the Evaluation of XML Retrieval », 2008. `http://www.inex.otago.ac.nz/`.

Lotan-Kochan E., « Analysing Graphic-Based Electronic Discussions : Evaluation of Students' Activity on Digalo », *in* , Springer (ed.), *EC-TEL 2006 : First European Conference on Technology Enhanced Learning*, Crete, Greece, p. 652-659, October, 2006.

Lund K., Prudhomme G., Cassier J.-L., « Using analysis of computer-mediated synchronous interactions to understand co-designers' activities and reasoning », *Proceedings of the International Conference On Engineering Design*, Cité des Sciences et de l'Industrie, Paris, France, august, 2007.

Minsky M., A Framework for Representing Knowledge, Technical Report n° MIT-AI Laboratory Memo 306, MIT, Boston, June, 1974.

W3C, « XQuery 1.1, W3C Working Draft 3 December 2008 », 2008. `http://www.w3.org/TR/2008/WD-xquery-11-20081203/`.

Wilkinson R., « Effective Retrieval of Structured Documents », *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, p. 311-317, July, 1994.