Modélisation et extraction de schémas dialogiques dans les traces d'interaction langagières des forges logicielles

J-P Sansonnet[1], S. Bonneaud[1], Gabriel Ripoche[1,2]
[1] LIMSI-CNRS, AMI Group, 91403 Orsay Cedex, France
[2] GSLIS, University of Illinois at Urbana-Champaign, USA

Résumé
Dans cet article nous présenterons un modèle cognitif, associé à son langage formel, pour la *description informatique*, en vue de l'extraction automatique dans les flots d'interaction langagières des collectifs médiatisés de séquences dialogiques particulières : les schémas dialogiques liés à l'organisation coopérative des tâches. Notre modèle/langage s'appuie sur l'analyse qualitative et quantitative du corpus des interactions langagières de la forge logicielle Bugzilla. Dans ce résumé, nous mettons l'accent sur les motivations de cette étude.

1. Contexte

Ce travail se situe dans le contexte d'une collaboration internationale en cours avec l'université UIUC de l'Illinois (Champaign-Urbana) intitulée « Organizational Dynamics of Software Problems in the Open Source Software (OSS) Collectives » qui pose deux questions essentielles à la limite des méthodes de développement de logiciel d'une part et de la socio économie de la communication d'autre part : a) Comment les grandes communautés du logiciel libre gèrent-elles *collectivement* des flux *continus* de problèmes logiciels ? et b) Comment ces communautés (re)conçoivent-elles de façon *collective* et *continue* leur(s) logiciel(s) ? ». En effet ces communauté de développement se distinguent radicalement par leur mode d'interaction qui est géographiquement très distribué et qui a recours systématiquement aux outils de l'Internet.

2. Le corpus des interactions dans Bugzilla

Dans ce cadre, nous avons été amenés à nous intéresser spécifiquement à la communauté des développeurs de la Forge-OSS Bugzilla qui a pour rôle, entre autres, de concevoir et de développer, en mode continu et collectif, un environnement de débogage pour le moteur de navigation Internet Mozilla. Grossièrement, dans Bugzilla toutes les interactions passent par des formulaires web, entièrement publics, qui comportent deux parties :
- en haut du formulaire, une partie qui est qualifiée de formelle car contenant des données de type purement informatique, consultable et modifiable par les développeurs selon *des procédures de type social* c'est-à-dire que seule la bonne volonté (goodwill) des actants garantit le bon usage.
- en bas, une partie qui est qualifiée d'informelle car contenant sous forme de *commentaires langagiers* les explications (justification et raison des actions), les propositions d'action, les discussions argumentatives, les choix, les décisions …

Dans Bugzilla, un formulaire (une page web) est associé à chaque 'bug' ouvert (c'est-à-dire tout type de problème rencontré dans le fonctionnement de Mozilla et identifié collectivement comme tel) ; une fois le bug corrigé, le formulaire est marqué comme « clos » mais reste consultable, il peut même être rouvert. Actuellement, plus de 220 000 bugs on été répertoriées par plus de 1500 développeurs et environ 50 000 sont en cours de correction, alors que Mozilla est en fonction depuis 1998. Les formulaires contiennent en moyenne 10 commentaires langagiers, ayant quelques phrases chacun : nous disposons ainsi d'un corpus, mis sous XML par UIUC, qui contient plus de 6 000 000 de phrases en 'anglais-international'.

3. Exploration socio-informatique du corpus : du statistique vers le langagier

En vue des questions posées plus haut, les socio-informaticiens explorent le corpus d'interactions de Bugzilla comme étant d'une part la mémoire et d'autre par la seule apparence

disponible des processus collectifs distribués (PCD), inhérents à l'activité du collectif, dont même les actants n'ont pas forcément conscience, et qu'ils cherchent à *exhiber*. Dans des précédents travaux nous avons commencé par effectuer des études de nature statistique sur la partie formelle des bugs ; par exemple, en suivant « la durée de vie des bugs » en la corrélant aux états successifs, pour proposer des modèles prédictifs. Cependant, deux phénomènes sont apparus : a) nous sommes vite arrivés aux limites des modèles statistiques, trop peu précis pour la recherche de *comportements fins* et b) les sociologues ont suggéré que, nous citons, « *la majeure partie de la sémantique des interactions est contenue dans les commentaires langagier* ». D'où l'idée de lancer un projet d'exploration de cette modalité avec pour objectif de caractériser et si possible d'extraire par des méthodes automatiques les processus du collectif. Cet objectif est très/trop ambitieux et en tout cas de longue haleine, même si nous avons fait deux postulats principaux, extrêmement restrictifs du point de vue du traitement de la langue : a) en raison du langage métier très ciblé employé nous ne prétendons aucunement nos choix et nos résultats généralisables ; b) parmi les phénomènes linguistiques du corpus, nous ne nous intéressons qu'à une classe bien particulière, les *schémas dialogiques*, ou séquences d'énoncés qui gouvernent l'organisation coopérative des tâches à effectuer.

4. Travaux effectués

Dans ce cadre trois études on été menées :

a) nous avons procédé dans un premier à une analyse qualitative puis quantitative des actes de dialogues des énoncés du corpus débouchant sur une ontologie spécifique au domaine. Rappelons qu'il s'agit de « coopération pour l'organisation de tâches » ; du coup, à côté des actes de dialogue généraux, on peut déterminer des actes plus fins. Cette ontologie a été validée par des mesures et contrastée avec l'ontologie de référence DAMSL.

b) en même temps, s'appuyant sur les analyses qualitatives, nous avons proposé un modèle cognitif pour les schémas dialogiques du corpus (appelé BIPs pour Basic Interactional Processes) et nous en avons défini un langage formel permettant de décrire au plan syntaxique et sémantique ces schémas. Ce modèle et son langage feront l'objet de l'exposé.

c) enfin nous avons commencé à développer des outils pour l'extraction automatique de schéma dialogues décrits dans notre modèle.

5. Conclusion

Le corpus Bugzilla est un outil public, précieux pour l'étude et l'extraction des phénomènes dialogiques dans un domaine particulier, mais qui va devenir crucial, celui des flots d'interactions langagières dans les activités des collectifs médiés par l'Internet. Il va bientôt permettre plusieurs types d'étude, et pour ce qui nous concerne, nous l'espérons l'extraction automatique de comportements *marqués* dialogiquement.

6. Bibliographie abrégée

Bugzilla, (accessed: 9/10/2004) http://bugzilla.mozilla.org/.

M. S. Elliott, W. Scacchi, Communicating and Mitigating Conflict in Open Source Software Development Projects. Software Development Projects.

G. Ripoche, L. Gasser, Scalable Automatic Extraction of Process Models for Understanding F/OSS Bug Repair. 2003 Conference on Software & Systems Engineering and their applications (ICSSEA'03) CNAM, Paris, France, December 2-4, 2003

E. de Vries, K. Lund, M. Baker, Computer-Mediated Epistemic Dialogue: Explanation and Argumentation as Vehicles for Understanding Scientific Notions. The Journal of the Learning Sciences, 11(1), 63-103, Lawrence Erlbaum Associates, Inc. 2002

J. Allen, M. Core, Draft of DAMSL : Dialog Act Markup in Several Layers, rapport, University of Rochester, Dept. of Computer Science, 1997

M.J. Baker Forms of cooperation in dyadic problem-solving. RIA, 16(4-5):587-620. 2002