ANALYSES ET PARTAGES DE CORPUS DE DISCUSSIONS AVEC CALICO - LEÇONS TIRÉES D'UNE EXPÉRIENCE RÉCENTE

François-Marie Blondel

Laboratoire STEF, ENS Cachan

Emmanuel Giguet

GREYC, Université de Caen

Résumé : La plate-forme Calico a été conçue et mise en place pour faciliter le partage de corpus de discussions et d'outils pour les analyser au sein d'une communauté de chercheurs et de praticiens. Cette communication dresse un bilan des outils disponibles, des corpus déposés et utilisateurs enregistrés sur la plate-forme en juin 2011 et propose une analyse des utilisations pour la période qui s'étend de 2008 à 2011. Conçus à l'initiative des chercheurs ou à la demande des utilisateurs, les outils disponibles relèvent de deux catégories, ceux qui offrent des vues globales et synthétiques sur les discussions et ceux qui permettent la recherche et la visualisation de marqueurs extraits du lexique. Cette deuxième catégorie d'outils a été la plus utilisée, parfois de manière intensive pour certains utilisateurs qui en ont tiré profit pour visualiser des marqueurs propres à leurs analyses. Un tiers environ des 283 discussions déposées en juin 2011 ont été rendues publiques, pour la plupart issues d'un corpus partagé issu du projet Mulce. Près de la moitié des discussions déposées sont restées accessibles à leur seul déposant. La charge de la conversion des données a été invoquée par plusieurs utilisateurs comme un frein au dépôt de corpus, l'anonymisation restant un obstacle pour les rendre publics. Des évolutions des outils pour permettre le dépôt en continu et le suivi des discussions dans la durée sont présentées au terme de cette analyse.

1. Introduction

Dans le domaine de l'apprentissage en ligne, les données dont peuvent disposer les chercheurs sont souvent issues d'expériences menées dans des contextes authentiques de formation. Ces expériences, difficiles à reproduire, rendent plus compliquées les comparaisons de résultats de recherche. Partager ces expériences, leurs résultats et les données associées revêt ainsi un intérêt tout particulier pour la communauté des chercheurs de ce domaine (Chanier & Ciekanski, 2010).

A partir d'un corpus donné de données de communication ou d'interaction, une grande variété d'analyses peuvent être conduites. Cependant, pour un type de corpus donné, comme par exemple un forum ou une conversation synchrone, la plupart des analyses font souvent appel à un ensemble d'outils commun pour décrire et représenter le corpus en question. Partager ces outils d'analyse peut rendre plus aisées les comparaisons de situations parfois difficiles à représenter, si l'on ne dispose que des descriptions textuelles.

La plate-forme Calico a été conçue et mise en place pour faciliter le partage de corpus et d'outils au sein d'une communauté de chercheurs et de praticiens. Avec pour intention supplémentaire de placer les outils pour l'analyse dans le même environnement de travail que les données à analyser.

Avant de développer notre analyse de cette plate-forme et des ses utilisations, nous allons situer cette initiative par rapport à quelques projets similaires.

1.1. Deux exemples de partages

Plusieurs projets de partage de données issues d'expérimentations se sont développés récemment, en réponse à des attentes institutionnelles ou à des propositions de recherche. Nous en retiendrons ici deux exemples qui nous paraissent proches des questions de partage évoquées plus haut.

1.1.1. PSLC DataShop

PSLC DataShop est d'abord un entrepôt qui centralise les données d'interaction issues de plusieurs dispositifs de cours en ligne, provenant pour la plupart des tuteurs virtuels du Pittsburgh Science of Learning Center. Les domaines d'enseignement couverts sont étendus, des mathématiques aux langues en passant par la chimie et la physique. Les données de base de l'interaction relèvent toutes d'un même modèle d'interaction assez simple qui permet de représenter les interactions au niveau le plus élémentaire ; il comprend trois éléments structurés, le contexte du message, l'action de l'étudiant (ou du tuteur) et la réponse du tuteur à l'action de l'étudiant. Les données annexes recueillies en même temps que les expérimentations sont associées à ces données de base d'interaction, comme par exemples des questionnaires ou des entretiens, mais aussi les articles publiés ou les présentations des résultats qui ont été effectuées.

Mais *PSLC DataShop* propose aussi des outils qui produisent des représentations graphiques donnant des indications sur le fonctionnement des tuteurs virtuels, comme des profils de performance, des rapports sur les erreurs ou encore des courbes d'apprentissage (Koedinger et al., 2009).

En février 2011, *PSLC DataShop* recensait 246 ensembles de données issus de 50 projets, représentant environ 150 000 heures d'interaction. Ces chiffres donnent une idée du succès (quantitatif) de ce site de partage. Succès que l'on peut interpréter par le caractère assez simple des formes d'interaction qui sont enregistrées (le modèle de données peut être utilisé pour nombre d'environnements d'apprentissage) et par la communauté de point de vue des

utilisateurs, pour la plupart membres du même regroupement de chercheurs (Pittsburgh Science of Learning Center). Il faut noter également que le *PSLC DataShop* bénéficie d'un soutien financier et institutionnel important, notamment de la part de la National Science Foundation.

1.1.2. Mulce

Le projet *Mulce* (Multimodal Learning Corpus Exchange) a centré ses propositions sur le partage de corpus d'interaction dans un contexte d'apprentissage. Il a étudié une série de questions sur le format des données d'interaction, les droits, les rôles, la description du contexte et des scénarios. Un effort particulier a été porté sur la structuration des corpus, en cherchant à s'appuyer sur les normes et les standards disponibles. La notion de corpus distinguable a été introduite pour identifier un sous-ensemble d'un corpus global et associer ce sous-ensemble à une problématique ou une question traitée (Reffay et al., 2008). Une plate-forme de dépôt a été mise en place qui comportait en juin 2011, 6 corpus globaux d'apprentissage et 26 corpus distinguables.

On pourra noter que ce qui fait l'intérêt de cette proposition – structurer avec précision les données recueillies lors d'une expérimentation et leur associer des métadonnées en respectant les normes – est aussi un obstacle potentiel à son adoption par les chercheurs « ordinaires ». En effet, très peu de corpus sont actuellement disponibles dans un état qui permette une intégration facile sur la plate-forme de dépôt de *Mulce*. Cette situation évolue lentement car les concepteurs de dispositifs d'apprentissage en ligne n'ont pas encore pris en compte la nécessité d'un export structuré des données d'interaction dans des formats d'échange ouverts et reconnus.

1.2. Calico, origine et utilisations

La proposition à l'origine de la plate-forme Calico se situe dans la lignée des deux initiatives précédentes. Il s'agit principalement d'offrir à la fois un espace pour entreposer des corpus d'interaction et des outils qui en facilitent l'analyse. Une catégorie de corpus est privilégiée, celles des discussions issues des dispositifs de communication textuelle.
Cette plate-forme[1] a été mise en place dans le cadre de l'ERTé Calico[2] (Bruillard, 2008) à destination des équipes participant à cette action de recherche sur l'emploi des forums dans des contextes de formation.

L'intention qui a guidé la conception de cette plate-forme était de fournir aux chercheurs des instruments qui facilitent la confrontation et les comparaisons de méthodes et d'outils d'analyses de forums de discussion. Avec une visée à plus long terme de transformer et d'affiner ces instruments en vue de leur utilisation par des formateurs et des tuteurs.
Ouverte en 2008 aux participants de l'ERTé, puis à un public plus large au milieu de l'année 2009, Calico est une des rares plates-formes de partage de corpus de discussions et d'outils en accès libre.

Dans cette communication, nous nous proposons de faire une analyse des utilisations de cette plate-forme afin d'en dégager les principaux usages et si possible de mettre en évidence les conditions et les facteurs qui rendent plus facile ou plus difficile le partage de corpus et d'outils.

Cette analyse a été effectuée en dressant un bilan de la plate-forme en juin 2011 (outils disponibles, corpus déposés, utilisateurs enregistrés) et des données disponibles sur les utilisations pour la période qui s'étend de 2008 à 2011.

[1] http://woops.crashdump.net/calico/
[2] http://www.stef.ens-cachan.fr/calico/calico.htm

1.3. Un format commun pour échanger différentes formes de discussions

Le format d'échange, *XMLForum*, utilisé pour importer, représenter et exporter des discussions, a été défini pour être le plus simple possible et permettre le dépôt et la diffusion du contenu de forums ou de tout autre forme de discussion de structure similaire, et ce, dans toutes les langues.

Dans sa forme la plus réduite, il considère une discussion comme une suite de messages caractérisés par une date, un auteur, un sujet, une référence éventuelle à un message précédent, et un contenu.

```
<message id="7">
    <header>
        <datetime>2006-10-07 11:43</datetime>
        <author id="5"> M.A.</author>
        <subject>Et il n'est pas toujours évident de communiquer...</subject>
        <msgref id="6"/>
    </header>
    <body>
    <content>
    Je suis d'accord avec toi sur le fait qu'on voit toujours les même au CDI et je
    m'aperçois (en collège) que les élèves, bien qu'ayant reçu une formation en
    6ème, oublient peu à peu qu'ils pourraient y trouver de nombreuses ressources
    pouvant les aider. Je pense [.....]
    </content>
    </body>
</message>
```

Figure 1 – Exemple de message au format *XMLForum* extrait d'un forum de discussion entre professeurs de documentation

Cette structure permet de représenter le contenu de forums, de listes de diffusion, mais aussi de chats ou de blogs.

2. Des outils pour des vues globales

Une première collection d'outils vise à offrir des vues globales sur une discussion indépendamment de sa taille, afin d'en donner une vision qui n'oblige pas à la lecture de tous les messages (*Anagora*, *Volagora*, *Themagora*).

Ces outils partagent quelques caractéristiques communes. La plus remarquable, inscrite dès les premières versions de ces outils, est de faire en sorte que les vues qu'ils produisent soient toujours lisibles quelle que soit la taille de la discussion analysée.

2.1. Volagora : volumes et temps

Cet outil a pour fonction de fournir des représentations graphiques des volumes et de leur répartition par périodes ou par auteur. La représentation par défaut affiche le nombre de messages par période sur toute la durée du forum.

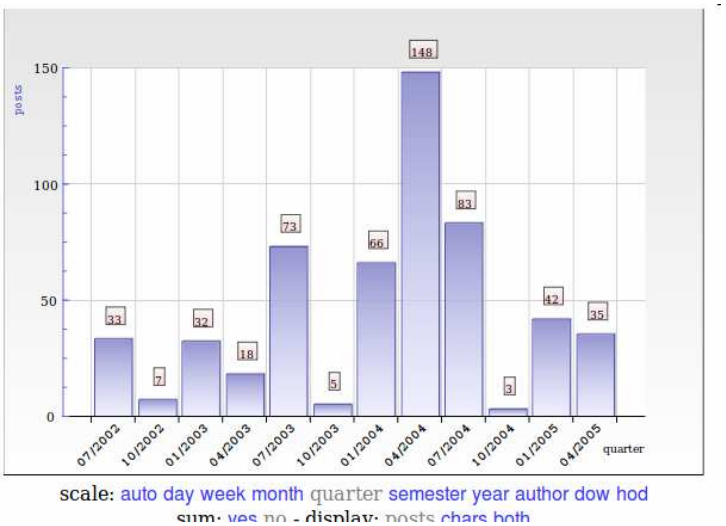

Figure 2 – Exemple de visualisation du nombre de messages par trimestre

Plusieurs autres représentations affichent les volumes (en nombre de messages ou de caractères) suivant un découpage temporel par jour, semaine, mois, .., année, ou encore par heures du jour ou jours de la semaine.

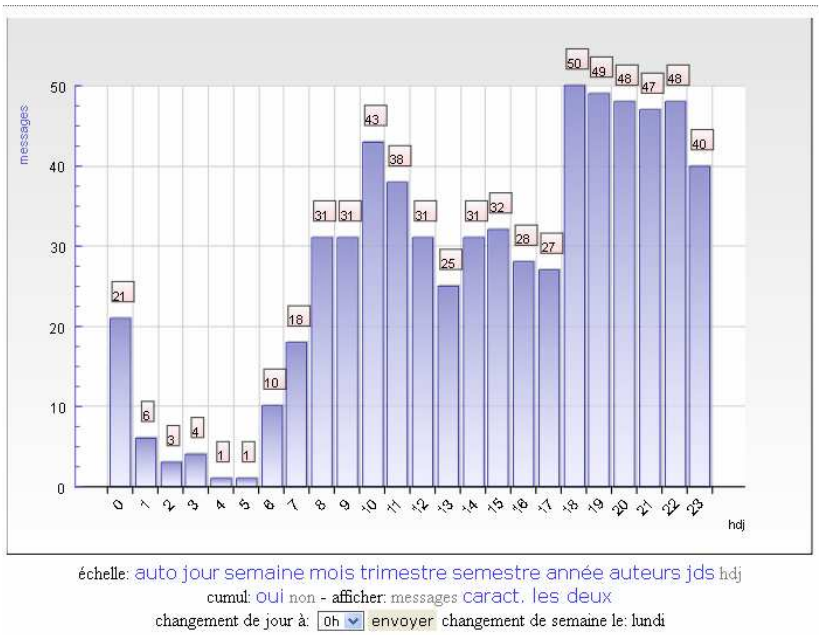

Figure 3 – Visualisation par heures du jour (permettant de constater que la période changement de jour dans cette discussion se situe entre 4h et 5h le matin)

Volagora affiche aussi le nombre de messages ou de caractères émis par un auteur, la distribution de la longueur des messages et celle des fils de discussion (en nombre de messages). Ainsi l'exemple de la figure 4 permet de constater que 23 messages n'ont pas reçu de réponse mais que 7 messages ont reçu de 6 à 10 réponses.

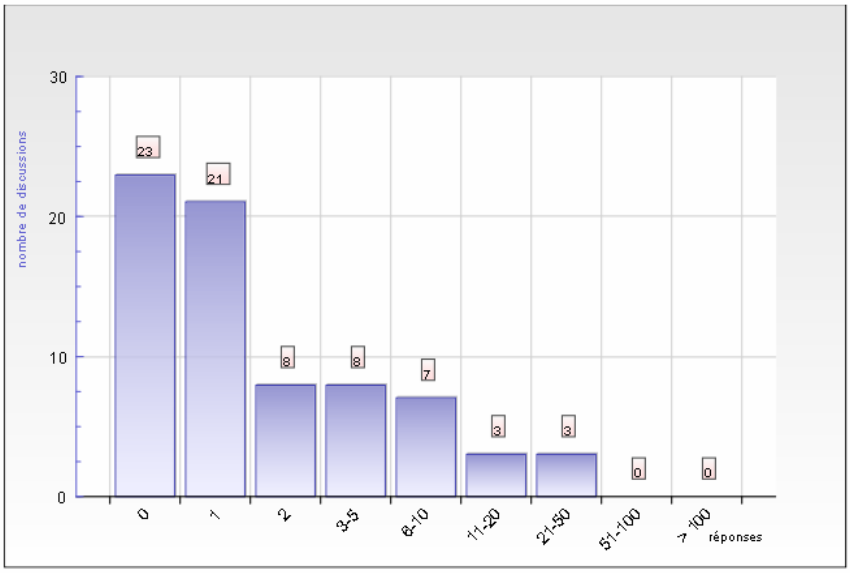

Figure 4 – Distribution du nombre de fils en fonction du nombre de réponses au premier message

2.2. Anagora : vues chronologiques

Anagora est un outil de visualisation des fils de discussion simultanés dans une même tranche temporelle. Il s'appuie sur le calcul du nombre de messages, et représente l'animation d'un forum.

Un fil de discussion est représenté par un bloc proportionnel au volume de messages échangés. S'il y a plus d'un fil de discussion au même moment, alors un nouveau couloir est dessiné au-dessus du premier. Chaque étage contient des fils de discussion distincts, dont le titre apparaît au passage de la souris. Ces étages, plus ou moins nombreux, sont appelés chronogrammes.

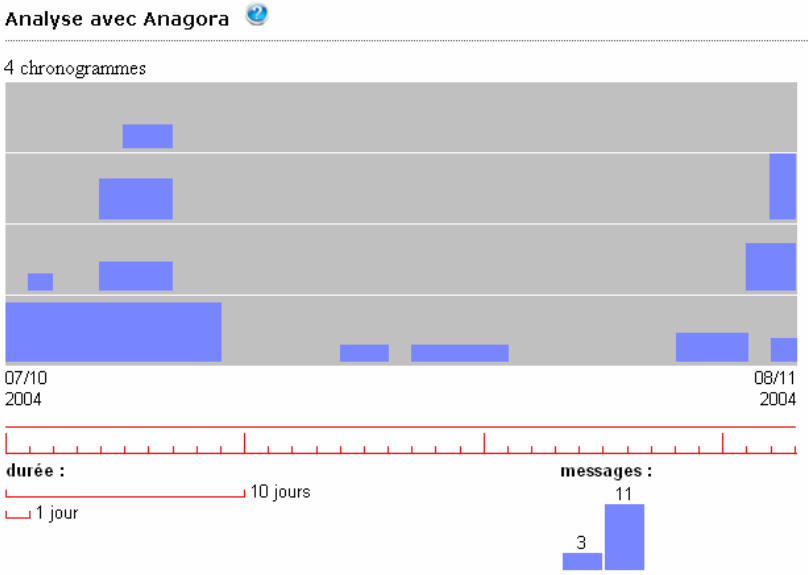

Figure 5 – Chronogrammes d'un forum d'un groupe de travail d'étudiants

L'exemple de la figure 5 montre que les échanges en parallèle ont lieu au début (discussions sur l'organisation du travail) et à la fin (sur le rendu du travail). Mais les fils de discussion simultanés peuvent être signe de désaccord s'ils interviennent au milieu du forum

et ne sont pas repris en fin, comme dans l'exemple suivant où les participants se disputent et échouent à fournir un travail commun (Figure 6).

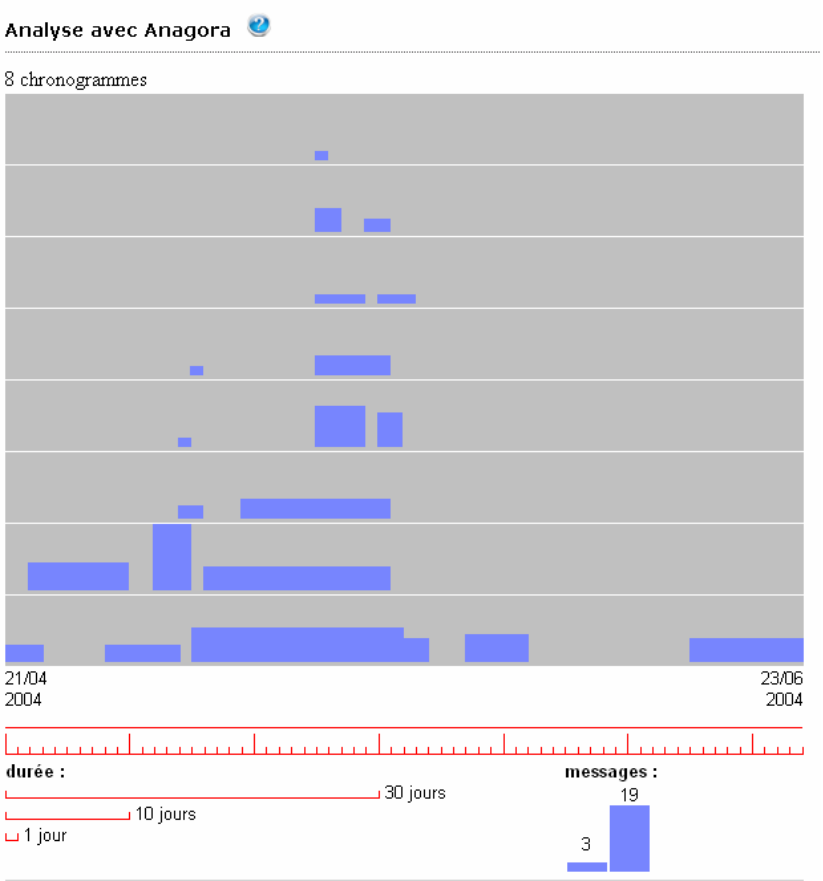

Figure 6 – Discussions simultanées nombreuses au milieu du forum

2.3. Authagora : auteurs

Cet outil liste une catégorisation des auteurs qui ont écrit des messages (contributeurs), initié des fils de discussion (initiateurs), écrit des messages sans réponse ou a contrario répondu à d'autres messages.

Analyse avec Authagora

33 contributeurs	6 initiateurs	3 auteurs sans réponse	31 auteurs qui répondent
M.ü (10)	A.M (2)	T.U (2)	M.ü (10)
C.é (3)	T.U (2)	A.M (1)	C.é (3)
A.M (3)	T.H* (1)	C.A (1)	A.T (2)
T.U (3)	Y.A (1)		D.O (2)
M-L (3)	C.A* (1)		N.H (2)
D.O (2)	M-L (1)		M-L (2)
N.H (2)			A.M* (1)
A.T (2)			C.I* (1)
Y.A (2)			Y.A* (1)
F.G (1)			G.A* (1)
Afficher tous les auteurs			

Figure 7 – Exemple de répartition des auteurs dans les quatre catégories

2.4. Themagora : une vue synthétique sur le forum comme un discours collectif

Themagora considère les messages successifs du forum comme un discours collectif sur lequel il va appliquer des règles d'analyse automatique pour y mettre en évidence des ensembles chronologiquement cohérents de messages appelés « moments ».
L'objectif est de proposer une vue synthétique des forums, pour permettre un diagnostic au temps t, notamment pour juger si la discussion est active ou au contraire languissante, si le tuteur doit intervenir ou non. Au départ conçu pour mettre en valeur l'exposition dans le discours collectif, le logiciel *Themagora* était destiné aux forums orientés par une tâche (Figures 8 et 9), spécialement les études de cas ou les devoirs. Il a été ensuite adapté aux forums libres (Lucas et Giguet, 2008).

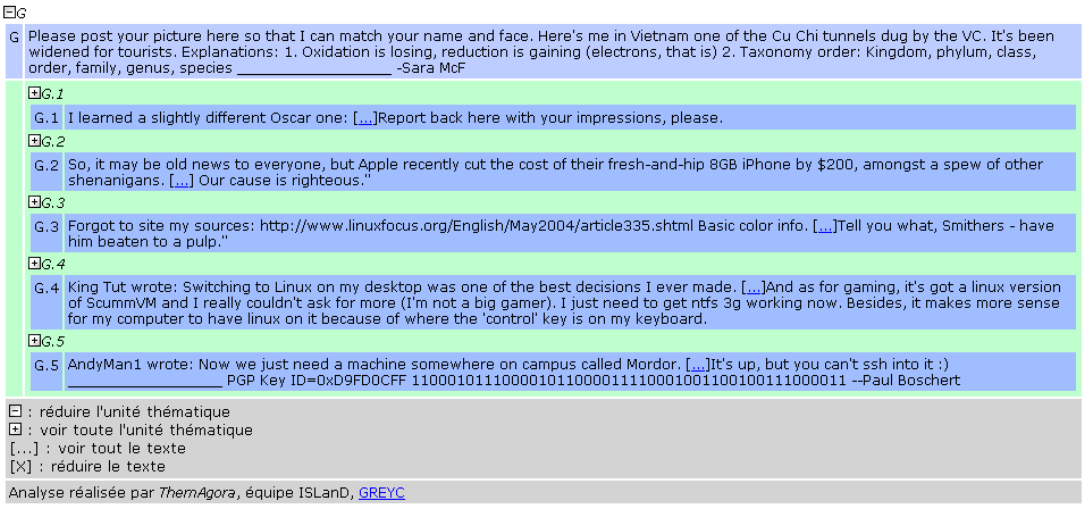

Figure 8 – Exemple de représentation d'un forum associé à une tâche (en anglais)

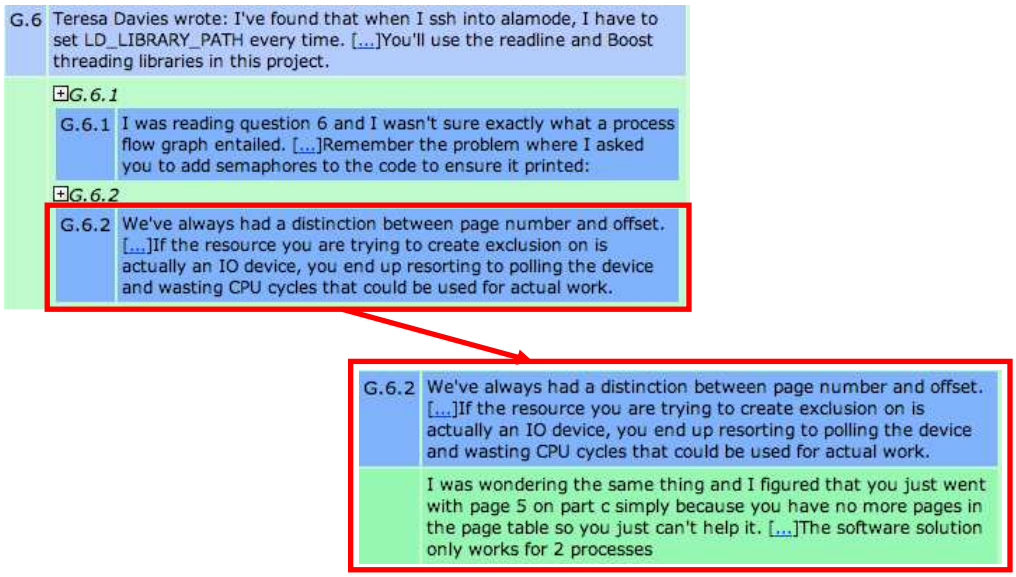

Figure 9 – Chaque partie peut être développée pour visualiser une plus grande quantité de texte

2.5. Des vues globales sur la totalité ou une partie des discussions

L'observation des représentations fournies par les outils présentés ci-dessous peut orienter les analyses vers certains sous ensembles particuliers d'une discussion. Les représentations peuvent ainsi être limitées par une période, ou par certains fils de discussion ou encore par

certains auteurs. Ces possibilités de restreindre la vue des affichages est disponible sous la description de la discussion (Figure 10) avec aussi une anonymisation des noms des auteurs.

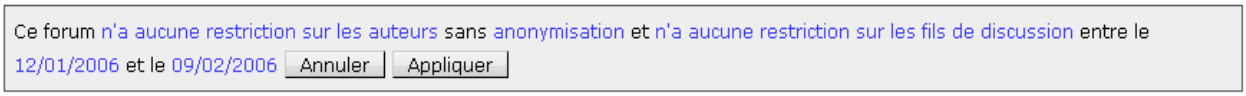

Figure 10 – Options pour restreindre les affichages à certains sous ensembles d'une discussion

3. Des outils pour lire, extraire, marquer et visualiser

Cette deuxième collection d'outils vise à faciliter la recherche de marqueurs voire d'indicateurs (Dimitracopoulou et Bruillard, 2007) par l'extraction du lexique et la définition de thématiques à partir de ce lexique, le comptage et la coloration thématique des messages (*Colagora*) et la visualisation de la présence de ces thématiques sur la totalité de la discussion ou suivant les fils de discussion ou les dates (*Bobinette*).
Tous ces outils fonctionnent indépendamment de la langue utilisée dans les discussions.

3.1. Colagora : construire le lexique et colorer les messages

L'outil *Colagora* qui fonctionne avec l'affichage du contenu des messages, a pour fonction de lister le lexique de toutes les formes employées et de permettre l'extraction, à partir de ce lexique, de formes regroupées dans un même thème d'étude.
Tout d'abord le calcul du lexique renseigne sur les occurrences des formes et sur les messages dans lesquels elles apparaissent. Chaque forme peut être utilisée directement ou ajoutée à un thème pour colorier les messages.

Sur l'exemple de la figure 11, le premier thème nommé *techno* regroupe des mots liés à des aspects techniques, le second nommé simplement *thème* regroupe des mots liés à la bibliothèque. Les mots de chaque thème apparaissent coloriés dans les messages qui sont affichés.

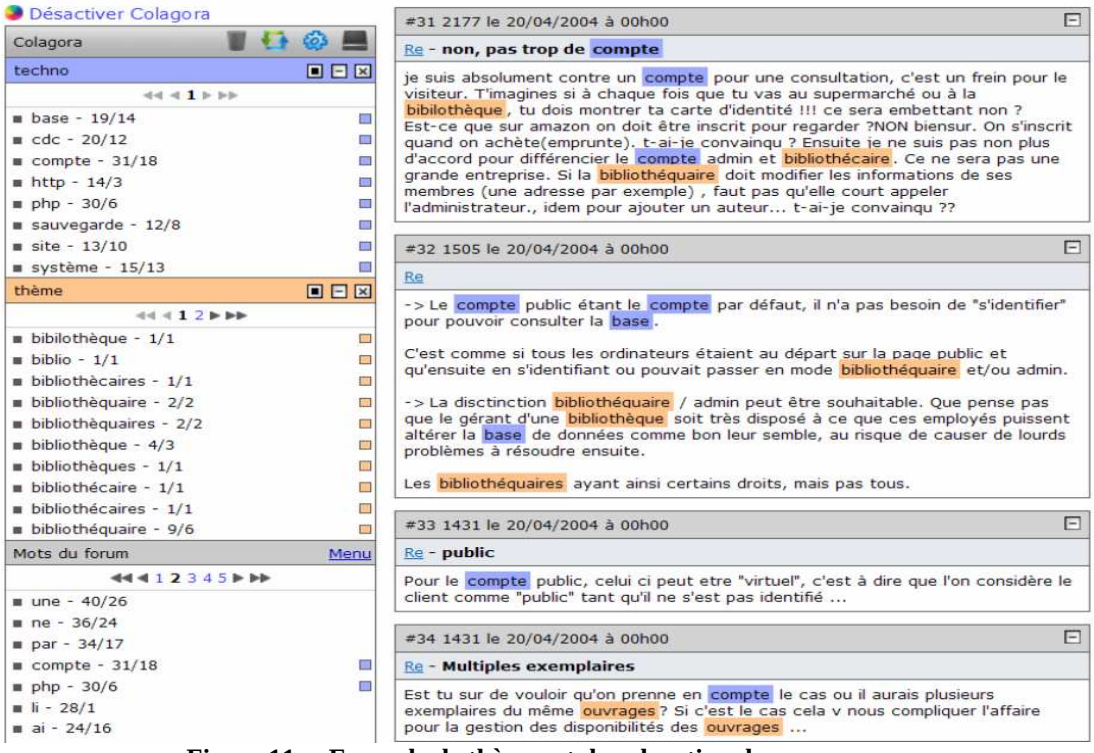

Figure 11 – Exemple de thèmes et de coloration de messages

3.2. Bobinette : suivre les fils et les thématiques

La première version de Bobinette a été conçue par Hyun Kim Bang et Bruillard (2005). Cet outil reprend l'idée d'une représentation de Nicole Clouet qui montre le déroulement chronologique des fils de discussion et qui affecte les messages d'une couleur ou d'un code suivant leur contenu.

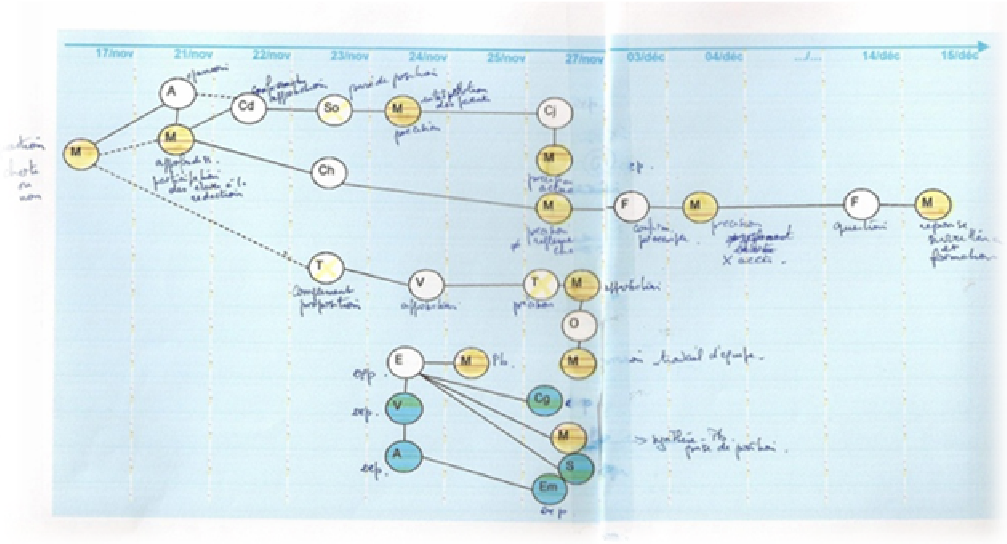

Figure 12 – la représentation initiale à l'origine de Bobinette

Cet outil a été ré-écrit pour être intégré dans Calico et permet de visualiser les fils de discussion des forums et de colorier automatiquement les messages avec les couleurs et les thèmes choisis dans Colagora.

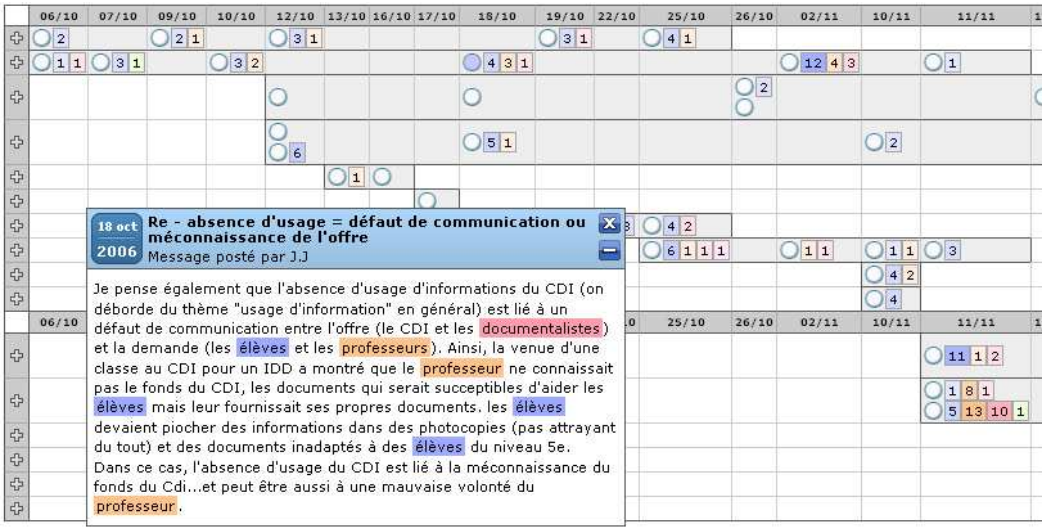

Figure 13 – Vue partielle de l'application de Bobinette sur une discussion avec quatre thèmes et affichage d'un des messages en surimpression

Chaque message est représenté par un cercle suivi du nombre d'occurrences des formes de chaque thème, avec une coloration d'autant plus saturée que l'occurrence est élevée. Le contenu des messages peut être visualisé au dessus de la représentation tabulaire.
Les thématiques qui peuvent être étudiées avec Bobinette sont très variées. La capacité de traiter toutes les langues est un atout pour l'analyse de discussions en formation en langues.

L'exemple de la figure 14 montre l'utilisation de l'anglais et du français dans 6 fils de discussion d'un groupe d'étudiants en apprentissage du français.

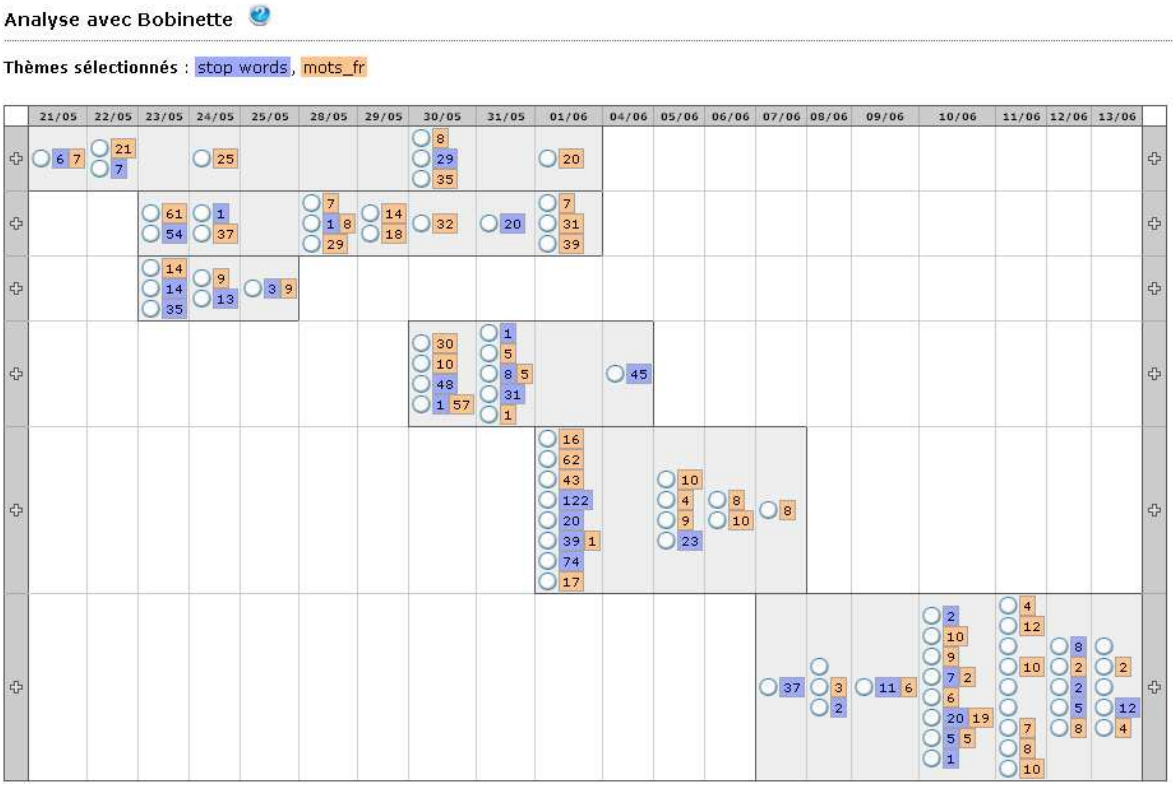

Figure 14 – La plupart des messages sont soit en anglais soit en français ; quelques messages bilingues apparaissent dans le dernier fil de discussion

L'outil Bobinette est plus adapté aux forums courts, car il est difficile de visualiser d'un seul coup d'œil toute la table d'un forum volumineux.

4. Utilisateurs et utilisations

Depuis le milieu de l'année 2009, la plate-forme est accessible à tous les chercheurs qui en font la demande. Chaque utilisateur identifié peut ainsi échanger avec les membres ou travailler sur ses propres corpus, préservant ainsi l'intégrité et l'anonymat de ces données.
Dans cette section, nous présentons un bilan des premières utilisations de la plate-forme en mettant l'accent sur deux aspects : les utilisateurs identifiés et les corpus qu'ils ont déposés d'une part et l'appropriation des outils par tous les utilisateurs d'autre part. Les données disponibles sont constituées du contenu de la base de données de Calico et les fichiers de journalisation de la plate-forme.

4.1. Utilisateurs, corpus et thématiques

En juin 2011, le nombre d'utilisateurs enregistrés s'élève à 53. Parmi ceux-ci, 17, soit environ un tiers, sont des membres des équipes de recherche de l'ERTé Calico. Les deux autres tiers peuvent être considérés comme des participants extérieurs même si quelques uns ont des relations de travail régulières avec les membres de Calico.

Si l'on s'en tient aux seules adresses de courriel des utilisateurs, 8 d'entre eux proviennent de pays non francophones : Grèce, Chine, Israël, Corée du Sud, Vietnam.

4.1.1. Discussions

Au total, 283 discussions ont été déposées depuis l'ouverture en 2008. Les discussions en accès public représentent environ un tiers du total (100 discussions) ; les discussions partagées au sein du groupe ERTé un huitième (35 discussions) ; la majorité des discussions (148) sont restées accessibles aux seuls déposants.

Publiques	Calico	Privées	Total
100	35	148	283

Tableau 1 – Discussions déposées depuis l'ouverture

Les déposants, au nombre de 17, représentent environ un tiers des inscrits. Hormis deux utilisateurs particuliers ayant déposé respectivement 91 et 79 discussions, le dépôt moyen par utilisateur est assez faible, la moitié des dépôts ne dépassant pas les 5 discussions.

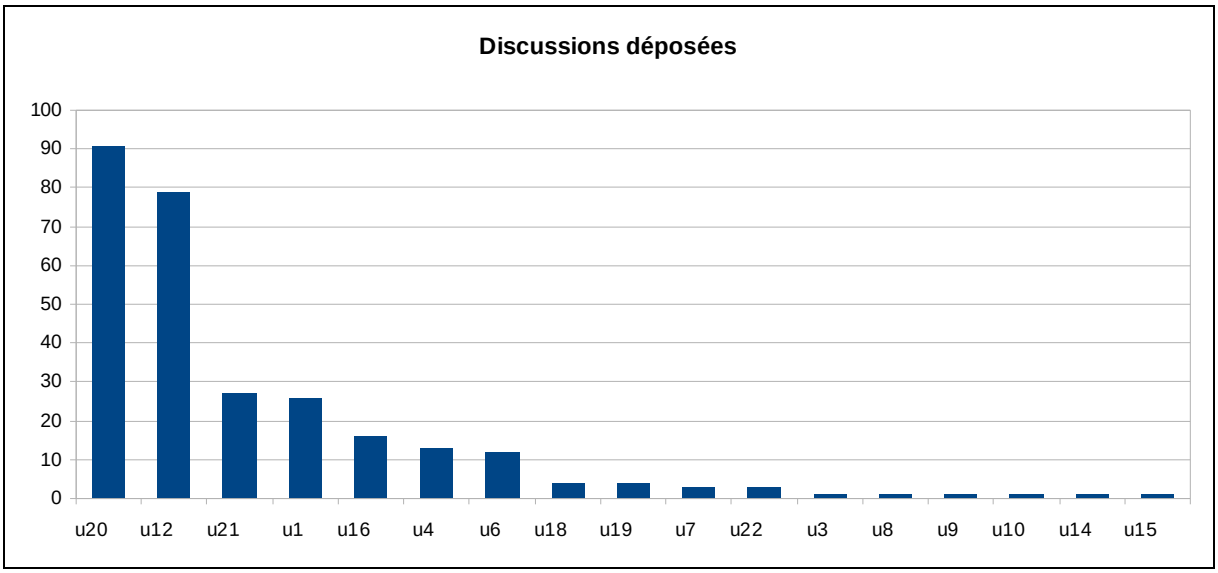

Figure 15 – Répartition du nombre de discussions déposées par utilisateur

Il est intéressant de noter que les 91 dépôts de l'utilisateur *u20* sont tous publics et que les 79 dépôts de l'utilisateur *u12* sont tous privés. La distinction public/privé est moins nette pour les autres déposants.

Un examen des quelques 135 discussions publiques ou accessibles au groupe Calico permet de constater la grande variété des origines et des formes de discussion : notes et commentaires (BSCW), discussions thématiques (forum débat), projets d'étudiants, cours en ligne, discussions libres (forums café), listes de diffusion (enseignants), blogs .La majorité des corpus déposés est issue de dispositifs d'apprentissage ou de formation en ligne.

Les discussions déposées sont de volume variable, de quelques dizaines à plus d'un millier de messages. En ce qui concerne le multilinguisme, des discussions ont été déposés dans cinq langues (français, anglais, grec, hébreu, vietnamien) et des analyses ont été faites en français, en anglais et en hébreu.

4.1.2. Thématiques

En juin 2011, le nombre de thématiques enregistrées s'élève à 105, dont à peine un tiers ont été rendues publiques.

Les créateurs de thématiques sont à peu près aussi nombreux (18) que les déposants mais 5 d'entre eux n'ont pas déposé de discussions, ils ont créé des thématiques pour analyser des discussions déposées par d'autres.

Hormis deux ou trois utilisateurs, la grande majorité a créé moins de dix thématiques et la moitié moins de cinq. A l'exception de *u12*, ceux qui ont déposé le plus de discussions ne sont pas ceux qui ont créé le plus de thématiques. On observe ainsi une sorte de partage (relatif) des tâches entre ceux qui sont majoritairement déposants et ceux qui sont majoritairement créateurs de thématiques.

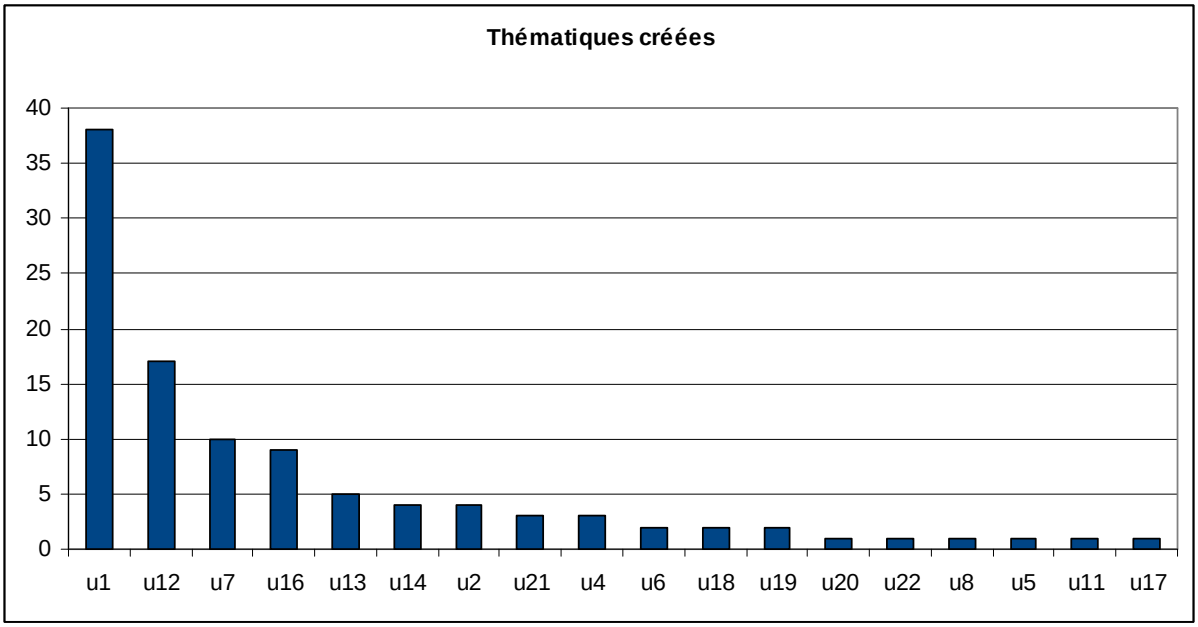

Figure 16 – Répartition du nombre de thématiques créées

4.2. Sessions

Une analyse des durées des sessions enregistrées sur la période février 2009 – octobre 2010, fait apparaître une grande variabilité (Figure 17). Le mode se situe sur une durée de 1 à 2 heures et représente presque le quart des 600 sessions dénombrées. Les sessions courtes, de moins de 5 minutes, en représentent un autre quart, tandis que les plus longues, de plus de 2 heures, représentent tout de même un dixième du total.

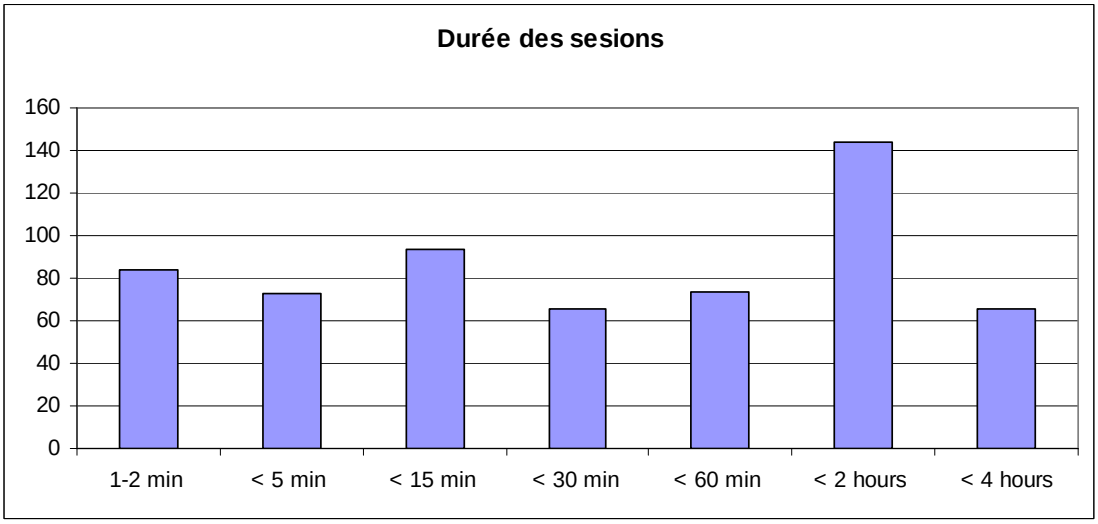

Figure 17 – Distribution des sessions de travail en fonction de leur durée

Si l'on comprend bien que certains utilisateurs n'ont fait que tester ou consulter les outils ou les fonctionnalités de la plate-forme dans des sessions courtes, la présence significative des

sessions longues est plutôt la marque d'un travail approfondi. Ces sessions longues rendent compte de plus de la moitié du temps total passé par les utilisateurs sur la plate-forme.

4.3. Fréquences d'utilisations des outils

Les données de journalisation de la plate-forme ont été analysées sur deux périodes consécutives, de février à décembre 2009 et de janvier à octobre 2010.

Les utilisations de la première période sont environ le double de celles de la seconde, pour un total de 2925 utilisations. En pratique, l'ouverture à un public de chercheurs dans le courant de l'année 2009 a donné lieu à de nombreux essais de la part des nouveaux utilisateurs et à des utilisations intensives de la part de quelques uns.

La répartition des utilisations sur les différents outils (Figure 18) s'explique par la manière de travailler avec la plate-forme.

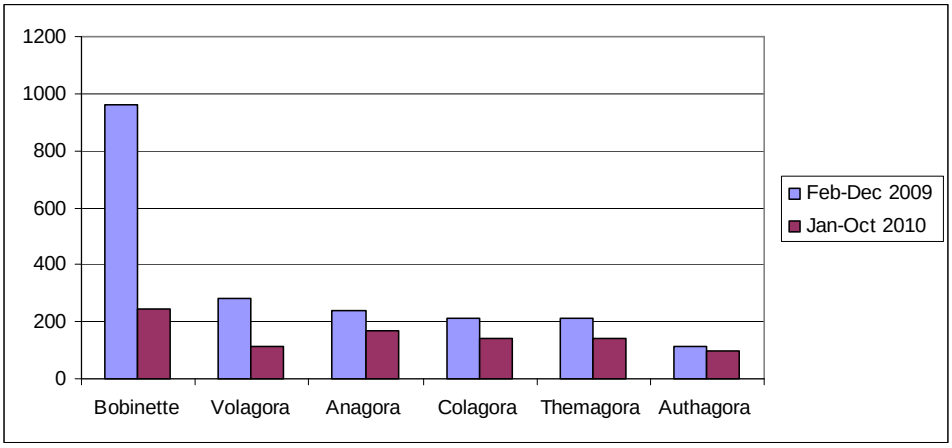

Figure 18 – Répartition des utilisations en fonction des outils par fréquences décroissantes

Pour une discussion donnée, les outils qui offrent des vues globales ne sont utilisés qu'une fois ou deux ; leurs résultats en changent pas d'une vue à l'autre, à l'exception de *Volagora* qui permet nombre de vues différentes. A contrario, l'outil *Bobinette* peut être utilisé un grand nombre de fois, au moins autant que le nombre de thématiques que l'utilisateur cherche à représenter, sans parler des visualisations de ces thématiques sur des périodes, des auteurs ou des fils particuliers.

4.4. Deux cas d'utilisation

Des exemples variés d'analyses thématiques ont été effectués (par exemple, les mots de la profession chez les stagiaires documentalistes définis par N. Clouet dans un travail antérieur). Au vu des discussions déposées et des publications liées à l'emploi de Calico, deux cas d'utilisation nous paraissent mériter une présentation particulière.

4.4.1. Simuligne

D'un point de vue quantitatif, la plus grande partie des discussions publiques (90 sur un total de 100) est constituée de forums de discussion issus de l'expérimentation *Simuligne*. Le fait que ces données aient été préalablement traitées et déposées sur la plate-forme de partage *Mulce* pour être rendues publiques a été un facteur particulièrement favorable à leur dépôt sur Calico.

Des analyses complémentaires sur ce corpus ancien ont pu être réalisées. Ainsi une recherche de marqueurs de cohésion dans un groupe à partir de l'emploi des pronoms a été utilisée en complément d'une analyse par les réseaux sociaux (Reffay & al., 2011).

4.4.2. Visualisation d'analyses multiples issues de Knowledge Forum

Le deuxième corpus important de discussions (79) provient d'un travail sur la construction collective de connaissances avec le dispositif *Knowledge Forum*.

Les données proviennent d'une expérimentation menée dans un établissement d'enseignement secondaire à Hong Kong. La recherche d'indicateurs, menée par une méthodologie propre à ce groupe de recherche, a porté sur l'argumentation, l'étayage et le contenu de discussions (Law & al., 2011). Plusieurs thématiques ont été créées pour repérer la trace de ces indicateurs dans les fils de discussion et l'outil *Bobinette* a été utilisé pour visualiser ces indicateurs dans la phase de recherche et dans la publication qui en découle (Law & al., 2011).

Les discussions et les thématiques associées à ce travail n'ont pas été rendues publiques. Ce travail de recherche de marqueurs pourrait expliquer en partie les utilisations plus importantes de la plate-forme et de ses outils dans l'année 2009.

A notre connaissance, il s'agit du premier travail d'envergure mené avec les outils de la plate-forme Calico en dehors du groupe de l'ERTé Calico.

5. Discussion

Les résultats de l'analyse des utilisations présentée ci-dessus pourraient orienter vers trois catégories principales d'utilisateurs.

Tout d'abord, les participants de l'ERTé Calico qui ont testé et utilisé la plate-forme au fur et à mesure de sa mise en place et de son développement. Une seconde catégorie d'utilisateurs « intensifs » a cherché à tirer parti des possibilités offertes soit pour partager des corpus soit pour exploiter les outils. Les deux cas particuliers présentés plus haut en sont des exemples caractéristiques. La troisième catégorie regroupe des utilisateurs plus « occasionnels » qui ont utilisé la plate-forme soit pour en observer le fonctionnement soit pour étudier une question particulière sur un nombre de discussions limité.

Bien que de nombreux chercheurs ou formateurs aient manifesté un intérêt pour les possibilités de partage offertes par Calico, il faut constater que plusieurs d'entre eux n'ont pas pour autant déposé de discussions ou proposé de nouvelles analyses des discussions déposées.

Lors des échanges que nous avons eus avec ces utilisateurs potentiels, plusieurs difficultés ont été soulevées notamment pour le dépôt de discussions ; la conversion des données depuis une plate-forme de formation vers le format d'échange de Calico est un obstacle technique dont la solution peut demander plus de temps que prévu. L'existence de connecteurs pour les plates-formes les plus employées en formation permettrait de lever cet obstacle,

Cependant, on peut aussi faire l'hypothèse de réticences au partage et à la publication de données comme l'a soulevé Nelson (2009). L'anonymisation, indispensable pour rendre publiques les données issues de communication entre personnes, est une entreprise qui peut s'avérer délicate à mener à terme. Il faut mentionner aussi que le travail supplémentaire que suppose la publication de données est relativement peu valorisé par les instances d'évaluation de la recherche, comparé à celui de la publication d'articles.

6. Perspectives concernant les outils

Afin de faciliter le dépôt de discussions, des techniques de dépôt en continu sont étudiées. Parmi celles-ci, l'abonnement à un flux RSS ou à une liste de diffusion sera bientôt proposé aux utilisateurs pour leur permettre d'alimenter automatiquement la plate-forme et ainsi d'analyser des discussions au fil de l'eau.

Dans cette perspective, il est aussi envisagé de transformer les techniques de traitement pour pouvoir prendre en compte des volumes de discussions plus importants, afin de permettre des études sur des temps longs. La possibilité d'afficher simultanément plusieurs vues sur une même discussion est aussi envisagée, dans le but de visualiser des différences et ainsi faciliter l'analyse des évolutions.

7. Bibliographie

Bruillard, É. (2008). Teacher development, discussion lists and forums: issues and results. In K. McFerrin, R. Weber, R. Carlsen & D.A. Willis (eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference, SITE 2008* . Chesapeake, USA: AACE. pp. 2950-2955.

Chanier, T. & Ciekanski, M. (2010). Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage, *Alsic*, vol. 13.

Clouet, N., Roué D., Bruillard É.. (2009). Forums for preservice teachers' development: Lessons learned from five years of research. In A. Dimitracopoulou, C. O'Malley, D. Suthers & P. Reimann (Eds.) *Computer Supported Collaborative Learning Practices: CSCL 09 Conference Proceedings* (pp. 214-218), International Society of the Learning Sciences (ISLS).

Dimitracopoulou, A., Bruillard E. (2007). Enrichir les interfaces de forums par la visualisation d'analyses automatiques des interactions et du contenu, *STICEF, vol. 13*, pp. 345- 397.

Giguet, E. et Lucas N. (2009). Creating discussion threads graphs with Anagora, In A. Dimitracopoulou, C. O'Malley, D. Suthers & P. Reimann (Eds.) *Computer Supported Collaborative Learning Practices: CSCL 09 Conference Proceedings* (pp. 616-620), Rhodes : ICLS.

Koedinger, K., Cunningham, K., Skogsholm A., & Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. In R.S.J.d. Baker, T. Barnes, & J. E. Beck, (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 157-166). Montreal, Canada.

Law, N., Yuen, J., Wong, O. W., & Leng, J. (2011). Understanding learners' knowledge building trajectory through visualizations of multiple automated analyses. In S. Puntambekar, G. Erkens & C. Hmelo-Silver (Eds.), *Analyzing Interactions in CSCL: Methodologies, Approaches and Issues* (pp. 47-82). Boston, MA: Springer.

Lucas, N. et Giguet E. (2008). Robust adaptive discourse parsing for e-learning fora. *The 8th IEEE International Conference on Advanced Learning Technologies (ICALT 2008),* Santander, Cantabria, Spain, IEEE Computer Society. pp. 730-732.

Nelson, B. (2009). Data sharing: Empty archives. *Nature, 461*(7261), 160-163.

Reffay, C. & Betbeder, M.-L. (2009). Sharing corpora and tools to improve interaction analysis. In U. Cress, V. Dimitrova & M. Specht (eds.) *Learning in the Synergy of Multiple Disciplines, Proceedings of the EC-TEL 2009, 4th European Conference on Technology Enhanced Learning*, LNCS 5794, Springer.

Reffay, C. Teplovs, C. & Blondel, F.-M. (2011).Productive re-use of CSCL data and analytic tools to provide a new perspective on group cohesion. In *9th International Conference on Computer-Supported Collaborative Learning, CSCL 2011*.

Reffay, C, Chanier, T., Noras, M. & Betbeder, M-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche, *STICEF*, vol. 15.

8. Sites Internet

ERTé Calico (2011). http://www.stef.ens-cachan.fr/calico/calico.htm
Plate-forme Calico (nd). http://woops.crashdump.net/calico/
Mulce plate-forme (nd). http://mulce.univ-bpclermont.fr:8080/PlateFormeMulce/VIEW/PUBLIC/01/accueil.jsp
Mulce-documentation (nd). http://ubpweb.univ-bpclermont.fr/HEBERGES/mulce/
PSLC Datashop (nd). https://pslcdatashop.web.cmu.edu/index.jsp
Knowledge Forum (nd). http://www.knowledgeforum.com/

Coordonnées des auteurs

François-Marie Blondel
Affiliation : STEF, ENS Cachan - Ifé, Cachan, France
Courriel : francois-marie.blondel@.ens-cachan.fr
Toile : http://www.stef.ens-cachan.fr/annur/blondel.htm
Adresse : Laboratoire STEF, ENS Cachan, Bâtiment Cournot, 61, avenue du Président Wilson 94235 Cachan cedex

Emmanuel Giguet
Affiliation : GREYC, Université de Caen, Caen, France
Courriel : emmanuel.giguet@info.unicaen.fr
Toile : http://users.info.unicaen.fr/~giguet/
Adresse : GREYC, Campus Côte de Nacre, Boulevard du Maréchal Juin, BP 5186 - 14032 Caen CEDEX