Symposium Corpus d'apprentissage en ligne : Conception, réutilisation, échange

Christophe Reffay, Thierry Chanier, Nikolaos Avouris, Laurent Romary

▶ **To cite this version:**

Christophe Reffay, Thierry Chanier, Nikolaos Avouris, Laurent Romary. Symposium Corpus d'apprentissage en ligne : Conception, réutilisation, échange. 2007. edutice-00161113

HAL Id: edutice-00161113
https://edutice.hal.science/edutice-00161113

Submitted on 9 Jul 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Symposium du colloque EPAL

Corpus d'apprentissage en ligne : Conception, réutilisation, échange.

Designing, re-using and exchanging online learner corpora

Le jeudi 7 juin 2007 à Grenoble, France

http://mulce.univ-fcomte.fr/epal_symposium/

*Pour améliorer la visibilité et la validité des travaux dans le domaine de l'apprentissage en ligne, pour permettre la discussion scientifique en profondeur sur les analyses, la comparaison des méthodes, des outils ou même des résultats, il manque bien souvent l'accès aux données du corpus initial. Le besoin d'échange de corpus existait dans d'autres domaines avant que nous l'exprimions comme un axe de méthodologie de recherche dans le domaine de l'apprentissage en ligne. Mais les objets considérés dans les collections (en TALN par exemple) sont bien moins complexes (tablettes, textes, livres, enregistrement de conversations téléphoniques, etc.) et plus homogènes. Construire un corpus issu d'une expérimentation d'apprentissage en ligne suggère de synchroniser et relier de nombreux fragments de données (traces d'accès, d'interactions, productions, tests, entretiens, etc.) pour qu'ils puissent faire sens dans une analyse des interactions a posteriori. Ce défi méthodologique a reçu récemment une reconnaissance scientifique à travers le projet Mulce (http://mulce.univ-fcomte.fr) par le soutien de l'Agence Nationale pour la Recherche. Pour progresser dans cette voie et susciter l'intérêt de cet axe méthodologique, nous organisons cet **atelier pratique** (symposium) invitant des chercheurs d'horizons divers pour ausculter les types de données recueillies, les outils et méthodes de traitement et d'analyse en vue de construire des corpus échangeables pour qu'ils puissent être traités par chacun.*

Symposium EPAL

Researchers in online learning are hampered in their efforts to improve the visibility and validity of their work, to promote in-depth discussion of their analyses, to compare methods, tools or even findings, by the fact that access to corpora in their original state is often impossible. In other fields, the need for corpora to be made available across different teams of researchers was expressed before we made it a research orientation in online learning. But in these domains, for example in ATNL, the artefacts in the data collections are less complex and more homogeneous (tablets, texts, books, recordings of telephone conversations etc) than in ours. Building a corpus from an online experiment involves synchronising and linking together many data fragments (log-on archives, traces of interactions, student products, tests, interviews etc) so that an analysis can be carried out post factum. This methodological challenge has recently gained scientific recognition, as is shown by the support of the Agence Nationale de la Recherche for the Mulce project (http://mulce.univ-fcomte.fr/). In order to further explore this area and to expose this methodological orientation to wider scrutiny, we offer a practical workshop (symposium). We are pleased to invite researchers from different horizons to discuss a wide range of issues such as types of data to be collected, tools and methods to be used in the processing and analysis of these data, such as might best contribute to the construction of corpora suitable for cross-disciplinary exchange and for treatment by different teams.

Référence de ce document :Reffay C. (2007). Symposium : Corpus d'apprentissage en ligne : Conception, réutilisation, échange. Colloque Echanger Pour Apprendre en Ligne (http://w3.u-grenoble3.fr/epal/), Grenoble, juin. http://mulce.univ-fcomte.fr/epal_symposium/

Responsable scientifique :

Christophe Reffay, Université de Franche-Comté, Informatique, Besançon, France.

Comité scientifique :

Nikolaos Avouris, Université de Patras, Human Computer Interaction, Patras, Grèce.

Marie-Laure Betbeder, Université de Franche-Comté, Informatique, Besançon, France.

Thierry Chanier, Université de Franche-Comté, Sciences du Langage, Besançon, France.

Christophe Reffay, Université de Franche-Comté, Informatique, Besançon, France.

Laurent Romary, Université Henri Poincaré, Informatique et TAL, Nancy I, France.

Comité d'organisation :

Marie-Laure Betbeder, Université de Franche-Comté, Informatique, Besançon, France.

Thierry Chanier, Université de Franche-Comté, Sciences du Langage, Besançon, France.

Maud Ciekansky, Université de Franche-Comté, Sciences du Langage, Besançon, France.

Elke Nissen, Université Stendhal, Sciences du Langage, Grenoble, France.

Muriel Noras, Université de Franche-Comté, Informatique, Besançon, France.

Christophe Reffay, Université de Franche-Comté, Informatique, Besançon, France.

Katerina Zourou, Université Stendhal, Sciences du Langage, Grenoble, France.

Symposium Epal

Intervenants :

Nikolaos Avouris, Université de Patras, Human Computer Interaction, Patras, Grèce.

Thierry Chanier, Université de Franche-Comté, Sciences du Langage, Besançon, France.

Christophe Reffay, Université de Franche-Comté, Informatique, Besançon, France.

Laurent Romary, Université Henri Poincaré, Informatique et TAL, Nancy I, France.

Philippe Teutsch, Université du Maine, Informatique, Le Mans, France.

Liste des participants :

Sara Alvarez
Isabelle Audras
Nikolaos Avouris
Camille Benabent
Marie-Laure Betbeder
Françoise Blin
Agnès Bracke
Eric Bruillard
Thierry Chanier
Maud Ciekanski
Christian Degache
Anne-Laure Foucher
David Gaveau
Liliana Gonzalez
Françoise Greffier
Christophe Jouis
Marie-Noëlle Lamy
François Mangenot
Sílvia Maria Martins Melo
Muriel Noras
Maguy Pothier
Christophe Reffay
Alessandra Rollo
Laurent Romary
Claude Springer
Philippe Teutsch
Simone Torsani
Vassileios VALMAS
Bruno Warin

Programme du symposium

Corpus d'apprentissage en ligne : Conception, réutilisation, échange.

Plénières – Amphi 3

09h00 – 9h30 : T. Chanier, C. Reffay
Introduction et présentation des enjeux de l'échange de corpus

09h30 – 10h00 : N. Avouris
Tools supporting Collaborative Learning and Interaction Analysis: Synergo and ActivityLens.

10h00 – 10h30 : L. Romary
Présentation des concepts fondamentaux de la "TEI " (Text Encoding Initiative) utilisés dans la "Freebank"

(pause)

Ateliers en parallèle (11h00 – 11h45)

- **Salle B101** - T. Chanier, C. Reffay, P. Teutsch
  Structuration, documentation, et parcours de « Simuligne » et « Copéas »
- **Salle B102** - N. Avouris
  Synergo and ActivityLens in action
- **Salle B103** - L. Romary
  La TEI en application

Symposium Epal

(11h45-12h30) Synthèse en plénière – Amphi 3

Chaque sous-groupe rendra compte en 5 minutes à l'assemblée, du contenu, de la structure, du contexte et du potentiel en analyse que le corpus visité recèle. Ce point de vue pourra être complété par l'intervenant ayant apporté le corpus en relevant les manques (soit les parties inexistantes actuellement dans son corpus, soit les parties existantes n'ayant pas été visitées lors de l'atelier). La synthèse de cette séance se fera sous forme d'une discussion générale ayant pour objectif de susciter l'intérêt sur les questions de l'échange de corpus.

(12h45-13h30) Repas – Salle « La Perouse / Coco » de la **Maison des Langues**

Un repas froid est prévu pour chacun des participants du symposium. Il sera servi à 12h45 dans la salle « Coco » de la Maison des langues.

Rappel : L'accueil et enregistrement pour le colloque EPAL se fait entre 12h à 14h.

Editorial

Parce que l'étude des dispositifs d'apprentissage en ligne traverse plusieurs disciplines, il y a différentes façons de définir la notion de corpus. Il est donc important que les partenaires du projet Mulce, à l'initiative de ce symposium, échangent avec une communauté plus large de chercheurs de différentes disciplines, leur vision de ce que peut être un « corpus échangeable » dans le domaine de l'apprentissage collaboratif en ligne. C'est donc dans un esprit d'ouverture et d'échange qu'a été conçu le symposium « Corpus ». Les trois interventions plénières proposées en début de ce symposium ont été choisies pour apporter des éclairages très différents sur cette notion de corpus, mais partagent sans conteste le besoin de rendre ces corpus échangeables.

La première intervention a pour objectif de planter le décor et d'exprimer les enjeux de l'échange de corpus. Elle est faite par Thierry Chanier (coordinateur du projet Mulce), dont l'objectif concret est de mettre en place une plateforme de partage de corpus issus de situation d'apprentissage collaboratif en ligne. Ayant déjà recueilli plusieurs ensembles de données, issus de différentes situations écologiques d'apprentissage des langues en ligne, les partenaires souhaitent à travers le projet Mulce, définir une structuration de ces données aussi complète et cohérente que possible, produisant un « corpus échangeable » pour permettre à d'autres équipes de les parcourir, les analyser, leur appliquer des traitements, et ainsi de construire au dessus de ces corpus, des analyses capitalisables.

Nikolaos Avouris et son équipe (Human Computer Interaction) de l'université de Patras ont une longue expérience de situations d'apprentissage collaboratif en ligne et sont les auteurs en particulier de Synergo : une plateforme d'apprentissage collaboratif, déjà utilisée dans plusieurs situations d'apprentissage au sein d'institutions différentes. Synergo a été finement paramétré pour tracer les différentes interactions des acteurs dans cet environnement lui permettant de les analyser pour construire un diagnostique ou un feedback à l'acteur. Nous aurons donc l'occasion ici d'observer des types différents d'interactions issues de domaines d'apprentissage autres que celui des langues. Les analyses qui en sont faites pourront également nourrir notre réflexion sur la structuration des corpus.

Enfin, Laurent Romary, dont les travaux sur la Freebank ont inspiré les objectifs de Mulce, revient sur les concepts fondamentaux de la TEI (Text Encoding Initiative) permettant de structurer des textes (ou transcriptions) en vue d'analyses lexicales, structurelles ou sémantiques. La Freebank rassemble des matériaux de base tels que des textes, des conversations audio par exemple ainsi que leur transcription pour permettre, très ouvertement, à d'autres chercheurs de disposer de ces ressources pour effectuer leurs propres traitements ou analyses. C'est la transcription initiale qui ouvre la porte à de nombreuses possibilités d'analyses. L'idée de la Freebank, reprise dans Mulce est donc de mettre à disposition des autres chercheurs, des données représentatives de contextes écologiques, dans un format permettant d'y appliquer de nouvelles analyses, elles-mêmes partageables, comparables et capitalisables.

Christophe Reffay

Support de cours

Documents joints : supports des intervenants

T. Chanier, C. Reffay. Extrait du projet ANR « Mulce », Juin 2006, adapté par C. Reffay pour le symposium EPAL\Corpus, Grenoble, 7 juin 2007. pp 5-24

N. Avouris, G. Kahrimanis, G. Fiotakis, E. Voyiatzaki, M. Margaritis. "Tools supporting Collaborative Learning and Interaction Analysis: Synergo and ActivityLens". Composition from previous publications of the HCI Group, Patras university, adapté pour le symposium EPAL\Corpus, Grenoble, 7 juin 2007. pp 25-49

S. Salmon-Alt, L. Romary, J.-M. Pierrel. "Un modèle générique d'organisation de corpus en ligne : application à la FreeBank", Traitement Automatique des Langues, Vol.45, n°3, pp. 145-169, 2004. pp 51-75

Référence de cet article : T. Chanier, C. Reffay. Extrait du projet ANR « Mulce », Juin 2006, adapté par C. Reffay pour le symposium EPAL\Corpus, Grenoble, 7 juin 2007.

Échange de corpus multimodaux d'apprentissage (Mulce)

Mulce (Échange de corpus d'apprentissage multimodaux, http://mulce.univ-fcomte.fr) est un projet soutenu par l'Agence nationale de la Recherche (ANR-06-CORP-006) dans le cadre du programme "Corpus et Outils de la Recherche en Sciences Humaines et Sociales". Il rassemble des équipes des laboratoires LASELDI et LIFC (Université de Franche-Comté), CREET (The Open University) et LIP6 (Université Paris 6), coordonnées respectivement par Thierry Chanier, Christophe Reffay, Marie-Noelle Lamy et Jean-Gabriel Ganascia.

1 Objectifs, contexte, problématique, originalité :

Etudier l'apprentissage en ligne, que cela soit à des fins de compréhension de cette forme d'apprentissage humain situé, d'évaluation des scénarios et dispositifs pédagogiques associés ou encore d'amélioration des environnements technologiques, requiert la disponibilité de données d'interaction provenant des différents acteurs, apprenants et formateurs, participant aux situations d'apprentissage.

Les publications et évènements scientifiques en rapport avec ce sujet ne manquent pas en France ou dans le monde. Mais les communautés pluridisciplinaires de chercheurs impliqués dans cette thématique n'ont pas encore réussi à caractériser un véritable objet d'étude scientifique, ni une démarche méthodologique en rapport. Les données sont inaccessibles à d'autres que les auteurs des écrits originaux. Elles sont parcellaires, donc décontextualisées, en regard des éléments constitutifs du dispositif de formation, ou encore inextricablement imbriquées au sein des environnements technologiques sous des formats propriétaires. Du coup le débat des chercheurs se déroule dans un espace où des conclusions contradictoires peuvent surgir sans que le jeu de la démarche scientifique ne soit véritablement convoqué. Souvent on cherche à comparer des objets aux contours mal définis, en fait différents. On ne peut réanalyser, répliquer, vérifier ni infirmer, étendre les résultats, toute chose pourtant à la base de la démarche scientifique.

Pour sortir de cette impasse, nous proposons la création et la diffusion de corpus d'un nouveau type, que nous appellerons "corpus d'apprentissage". Cet ensemble de données doit rassembler, non seulement les données résultats d'une formation mais également son contexte, c'est-à-dire les données caractérisant le dispositif de formation, ainsi que celles provenant du dispositif de recherche.

Ces données sont fortement multimodales à plusieurs titres très différents : les productions des participants peuvent utiliser des modes variés ; les nouveaux environnements synchrones ouvrent des espaces de production et de communication dans des modes et modalités interreliés ; enfin, les vidéogrammes issus des dispositifs de recherche, apportent un "fondu à plat" de ces espaces dont il convient d'extraire les constituants aux fins d'analyse. Avec la notion de corpus d'apprentissage apparaît donc la problématique des transcriptions, annotations, analyses multimodales, rencontrée dans d'autres champs des sciences humaines et sociales mais qui doit ici être repensée dans le cadre spécifique d'humains participant à des groupes d'apprentissage et interagissant dans des environnements technologiques appropriés.

La constitution de corpus d'apprentissage n'a d'intérêt que si ceux-ci peuvent faire l'objet d'échanges entre les communautés de chercheurs, ce qui oblige à : 1) les structurer et les formater suivant un modèle (à inventer) compatible avec les standards de corpus et de modèles pédagogiques déjà existants ; 2) les déposer sur un serveur répondant aux principes d'interopérabilité et d'accès libre ; 3) élaborer une charte éthique (car nous travaillons sur des productions d'individus), contrats de cession des droits et d'utilisation en rapport.

Mais faire du corpus d'apprentissage un objet d'études scientifiques nécessite aussi de le doter d'une méthodologie d'exploitation en rapport. C'est pourquoi une partie du projet Mulce

Symposium EPAL

s'intéresse à toute la chaîne de traitement, transcription, annotations, étiquetages, analyses, et aux outils associés. Les traitements successifs doivent bien sûr venir compléter chaque corpus d'origine.

Avec la mise à disposition de ces corpus, ainsi que les outils et services associés, s'ouvre alors la perspective de faire réanalyser des données par des équipes non initiatrices des projets de recherche ayant généré les corpus, ou bien de comparer des analyses de données prises dans des corpus différents, voire d'étalonner des outils de traitement ou d'analyse. Même si nous avons pris soin de construire le projet Mulce sur un ensemble de données et d'expertises maîtrisées par ses partenaires, on ne saurait prétendre amorcer une dynamique d'échanges au sein d'une communauté pluridisciplinaire et internationale sans y incorporer les premiers levains. On les retrouvera dans les tâches de partage prévues avec nos collaborateurs internationaux.

Contenus, objectifs et résultats attendus du projet Mulce

Le projet Mulce a les objectifs suivants :

1. Contribuer à créer au niveau de la communauté internationale et pluridisciplinaire des apprentissages en ligne une démarche de partage de données, de modèles, d'outils et d'analyses : Le symposium « Corpus » qui s'est déroulé le 7 juin à Grenoble, lors du colloque EPAL constitue une action concrète attachée à cet objectif et a plus précisément pour but :
   - de réunir une partie de cette communauté et de renforcer les liens ;
   - de préciser les axes scientifiques qui intéressent cette communauté ;
   - de susciter de nouvelles collaborations en vue de partager les corpus d'apprentissage en ligne, soit en ouvrant leurs corpus à d'autres, soit en proposant des outils qui permettent de les parcourir/traiter/analyser.

   Ces nouvelles collaborations sont appelées « collaborations extérieures » dans la suite du texte simplement pour les distinguer des partenaires « internes » du projet ANR Mulce lui-même.
2. Développer la notion de corpus d'apprentissage (différente de celle de corpus d'apprenants) afin qu'il devienne une base de références pour comparer, compléter les recherches sur les modèles, les outils et les analyses.
3. Développer un modèle (*Mulce-struct*) pour la structuration des corpus d'apprentissage qui permette l'échange de ces corpus à des fins de recherche et d'enseignement dans le respect des standards : ceux des corpus linguistiques (TEI, TEI Iso/TC 37/SC4) et, ceux orientés vers la conception pédagogique (IMS-LD).
4. Elaborer une charte éthique, des contrats de cession des droits et contrats d'utilisation à des fins d'enseignement et de recherche. Ces documents devant être validés par les collaborateurs extérieurs.
5. Installer un serveur Internet *Mulce-OAI* permettant la diffusion des corpus d'apprentissage (cf. ci-après) en accès libre, dans le respect des contrats précédents. Serveur validé OAI-MPH.
6. Déposer dans *Mulce-OAI* nos corpus d'apprentissage *Simuligne* et *Copéas*, mis au format *Mulce-struct*, après validation par nos partenaires.
7. Faire déposer dans *Mulce-OAI* par un collaborateur extérieur au moins un nouveau corpus d'apprentissage, mis au format *Mulce-struct*.
8. Adapter des outils d'analyses textuelles déjà développés par les partenaires Mulce au besoin de l'analyse de contenu (Valcke & Martens, 2006) des corpus déposés dans *Mulce-OAI.*

Symposium Epal

9. Faire déposer des analyses ou annotations effectuées sur les corpus *Simuligne* ou *Copéas* par les collaborateurs internes et extérieurs au projet. Ce travail correspond au niveau de description secondaire du corpus, qui devra être lié au premier niveau sans le modifier (voir la notion d'annotation externe versus interne (Salmon Alt, Romary & Pierrel, 2004) joint aux supports du symposium EPAL\Corpus).
10. Transcrire et modéliser les interactions multimodales, typiques de celles rencontrées dans les environnements audio-synchrones en ligne, interactions extraites du corpus *Copéas*.
11. Recenser les outils de transcription compatibles avec *Mulce-struct* et déposer notre outil de transcriptions multimodales, *Tasync*, dans *Mulce-OAI.*
12. Mettre en ligne dans *Mulce-OAI* des services de visualisation et de traitement des corpus compatibles *Mulce-struct*, notamment *Padis* (finalisé par Mulce), *Vicodili* (Lium) et ceux des autres collaborateurs extérieurs.
13. Organiser un colloque international centré sur la comparaison des méthodes et outils de transcription et d'analyse des interactions en ligne. Les organisateurs auront au préalable effectué des tests sur les corpus diffusés par *Mulce-OAI*.

On retrouvera chacun de ces résultats attendus dans la section 2, "Agenda" et les sections présentant les corpus, outils, analyses déjà existantes au sein des partenaires Mulce. Les tâches impliquant les collaborateurs dits "extérieurs" figurent dans l'agenda (section B2) et sont reprises en B4 (Collaborations internationales).

Nos atouts

Le projet Mulce présente les atouts suivants :

- Les partenaires du projet rassemblent des équipes pluridisciplinaires ayant une habitude éprouvée de travail en commun.
- Le projet dispose de corpus de très grandes valeurs sur le plan international du fait de leur taille, de la variété des données d'interaction provenant de situations d'apprentissage de nature écologique, intégrant les différents types d'environnements d'apprentissage en ligne, y compris les environnements synchrones multimodaux.
- Ces partenaires ont une expertise sur toute la chaîne de traitement associée à ce type de corpus : recueil et organisation des données, transcription et annotations, analyses manuelles et automatiques, développement d'outils de transcription, de visualisation, d'analyse.
- Ils participent aux réunions d'experts internationaux qui s'intéressent aux recueils et analyses des interactions provenant de situation d'apprentissage en ligne. Les partenaires Mulce ont inauguré à l'occasion du symposium EPAL\Corpus les premières rencontres scientifiques sur cette nouvelle thématique des corpus et ont à ce titre l'initiative au plan international.
- Les collaborations internationales associées au projet sont de grande qualité. Elles impliquent nos partenaires dits "extérieurs" à Mulce, qui disposent d'expertises complémentaires en différentes parties de la chaîne de traitement évoquée précédemment. Ces extérieurs sont déjà impliqués dans des collaborations avec les partenaires internes, sur les projets de recherche associés à nos corpus, ou sur des événements scientifiques en rapport. Leur participation, précisément caractérisée dans ce document, offre une perspective réaliste de voir l'échange de corpus d'apprentissage multimodaux ne pas se limiter au cadre restreint du partenariat interne, pour s'ouvrir à celui des communautés de recherche spécialisées présentes sur la Toile.

Symposium Epal

2 Description du projet et résultats attendus

2.1 Intérêt scientifique des corpus d'apprentissage

Les corpus en linguistique et TAL

Dans de nombreux domaines de sciences humaines, et particulièrement en linguistique (appliquée) et Traitement Automatique du Langage (TAL), la constitution de corpus de références, par la qualité des méthodes de recueil, la représentativité des données au regard d'un domaine circonscrit, sa structuration, est l'objet de nombreuses recherches et programmes (Daille & Romary, 2001 ;). Ils servent à la fois de données à partir desquelles les communautés effectuent des études sur le langage et les langues, de façon automatique ou manuelle (mais fréquemment assistées par des outils automatiques) et de base de références pour évaluer des outils de traitements automatiques concurrentiels ou complémentaires. Cela concerne aussi bien les corpus (et traitements) écrits qu'oraux (Blache et al., 2004), même si ces derniers sont encore peu nombreux (Véronis, 2004).

Les corpus sont donc devenus des objets partagés au sein de communautés de recherche. Pour cela ils doivent être **échangeables**, donc être organisés suivant des modèles de structures (pour les données et métadonnées) et des formats qui acquièrent le statut de standards, voire de normes. Ces points, bien qu'ayant retenu l'attention des chercheurs depuis les années 80 (citons, par exemple, la TEI (2006) et le CES (2000)) sont toujours l'objet de développements importants (voir groupe Iso TC37/SC4 (2006), réseau Atonet (2006), etc.). L'échange ne concerne, bien sûr, pas que les données recueillies (il serait alors unidirectionnel) mais aussi différentes étapes d'analyses, souvent dénommées **annotations**. Celles-ci font l'objet d'études dans deux perspectives complémentaires : chaque niveau d'analyse doit pouvoir se décrire suivant une structure qui vient compléter les données de base et des outils assistent les processus d'annotations (voir ce site, non maintenu, en guise d'aperçu des richesses d'approche (Cocosda, 2002)).

Dans le domaine de l'analyse des interactions orales, on peut citer au moins deux projets (CLAPI , 2006) et (TALKBANK, 2006) qui partagent certains objectifs de Mulce, mais sur des ensembles de données ciblés. Certaines de ces conversations peuvent être issues de situations d'apprentissage. Il s'agit alors par exemple de l'enregistrement audio/vidéo d'une session, d'un cours ou d'un atelier de travail. Les analyses concernant ces corpus s'intéressent à la structure et au contenu des conversations en alignant des transcriptions sur les matériaux (audio/vidéo). Mais ces deux projets proposent aussi des outils permettant d'aligner (loogiciel CLAN pour la Talkbank) , d'annoter, d'interroger ou de parcourir conjointement la transcription et le matériau source pour CLAPI. Ce dernier propose également l'interrogation de plusieurs corpus sur des phénomènes de recouvrement ou la recherche de mots (token) précis par exemple.

Ces travaux dans leur ensemble concernent des corpus plus restreints en général, mais leurs avancés sur la structuration, l'alignement, l'anonymisation et les règles éthiques, la diffusion, la mise à disposition d'outils de traitement ou de parcours guident nos propositions vers des objectifs similaires, mais avec des ensembles de données plus complexes.

Définition de la notion de corpus d'apprentissage

Notre domaine de recherche s'inscrit dans les sciences et technologies de l'information et de la communication pour l'éducation et la formation (aussi identifié sous le sigle TICE ou encore EIAH - environnements informatiques pour l'apprentissage humain -). C'est le cas en particulier du sous-domaine de l'apprentissage / formation en ligne, qui comprend les situations d'apprentissage survenant sur les réseaux informatiques, que cela soit en situation de formation à distance (FAD) ou dans tous les dispositifs associant présentiel et distance en proportion variable, le tout étant généralement étiqueté FOAD (formation ouverte et à

Symposium EPAL

distance). Les données résultats des actions des participants, des interactions entre participants ou entre ceux-ci et les systèmes sont les éléments fondamentaux à partir desquels opèrent les équipes des différentes disciplines concernées (sciences de l'éducation, du langage, psychologie, informatique, etc.). Elles servent à de multiples fins dont celles de comprendre l'apprentissage humain, à identifier les situations et dispositifs qui l'étayent, à renseigner les systèmes qui servent de médium, voire de médiation à ces apprentissages, à analyser les usages que font les humains des systèmes utilisés dans les dispositifs de formation. Elles peuvent également servir à l'évaluation des apprenants (connaissances, stratégies) ou des systèmes. Mais pour passer de données à un objet scientifiquement analysable, il faut ajouter encore d'autres données de qui constitueront par leur apport ce que nous appellerons un corpus d'apprentissage.

Plus précisément un **corpus d'apprentissage** est, **à un premier niveau (primaire)**, constitué de trois ensembles :

- Da) Les données résultats d'une formation : cela comprend aussi bien les productions des apprenants (produits des tâches programmées dans le dispositif de formation, réalisés seul ou en groupe), que les interactions des différents participants à la formation (comme les échanges synchrones ou asynchrones des apprenants, des enseignants / tuteurs, des experts, etc.), que (dans une forme moins élaborée) les traces (*logs.*) des actions, directes ou induites par l'architecture des systèmes, des acteurs dans les systèmes (connexions, traces d'actions d'un acteur dans telle partie de logiciels, échanges entre serveurs, etc.).
- Db) Les données extraites du dispositif de formation de façon à permettre l'identification de la situation d'apprentissage. Un certain nombre est extrait du dispositif tel que conçu avant la formation (scénario pédagogique, partie de matériaux pédagogiques, consignes, etc.). D'autres résultent de la mise en œuvre du dispositif et rendent compte des écarts avec la situation initialement planifiée (absences, évènements imprévus, apports de données extérieures, etc.).
- Dc) Les données provenant du dispositif de recherche : information sur les acteurs, questionnaires, entretiens, prise de notes, enregistrements vidéo, etc.

A un second niveau, le corpus d'apprentissage s'enrichira, comme dans les corpus cités en linguistique, des différents types de travaux de recherche subséquents à la formation : transcription, annotations, étiquetages, analyses corrélées. Ces **niveaux de description** sont donc en lien de dépendance avec le niveau primaire.

Circonscrire le domaine, les corpus d'apprenants

De notre définition, il ressort que, l'étude de l'utilisation de corpus à des fins pédagogiques, par exemple en classe de langue avec des concordanceurs (Tribble & Jones, 1997), ne font pas partie de notre objet d'étude et ne peuvent être dénommés corpus d'apprentissage, en particulier du fait que les textes de référence ne sont (en général) pas produits par des personnes en situation de formation. Il est par ailleurs évident que le mot apprentissage concerne l'apprentissage humain et qu'il n'est ici pas question de collecter et d'organiser des données avec des objectifs d'entraînement de systèmes automatiques. Le domaine *machine learning* est donc exclu.

D'autre part, ce qui est dénommé **corpus d'apprenants** (*learner corpora*) ne représente qu'un cas particulier et restreint de corpus d'apprentissage. En effet, ils regroupent uniquement des productions d'apprenants, pas celles d'autres acteurs. Ce sont souvent des produits récoltés en situation de test des connaissances et non en situation de formation (les interactions des acteurs ne sont alors pas présentes). Même si ces corpus d'apprenants peuvent être d'un grand intérêt pour les chercheurs utilisant des techniques de TAL en apprentissage des langues assisté par ordinateurs (Granger, Vandeveter & Hamel, 2001 ; Granger et al., 2002), ils ne permettent pas d'étudier des situations d'apprentissage. Ces

Symposium EPAL

collections de données peuvent bien sûr être accueillies dans le serveur Mulce, à condition d'être accompagnées de données/métadonnées sur le contexte, les acteurs (cf. données de types Db). Cela les rendrait ainsi accessibles à la communauté des chercheurs, ce qui n'est pas le cas aujourd'hui (cf. infra sur la question importante de l'accessibilité).

Intégrer le contexte, une exigence méthodologique

Un corpus d'apprentissage ne doit donc pas se limiter aux données provenant des apprenants. Il doit aussi incorporer les éléments définitoires du dispositif de formation (Db), complétés par celles du dispositif de recherche (Dc) pour constituer un objet digne d'études :

> *Researchers must carefully document the relationships among media choice, language usage, and communicative purpose, but they must also attend to the increasingly blurry line separating linguistic interaction and extralinguistic variables. […] Studies of linguistic interaction will likely need to account for a host of independent variable: the instructor's role as mediator, facilitator, or teacher; cross-cultural differences in communicative purpose and rhetorical structure; institutional convergence or divergence on defining course goals; and the affective responses of students involved in online language learning projects. (Kern, Ware & Warshauer, 2004).*

Notre domaine de recherche s'intéressant, non seulement à l'apprentissage mais également à la pédagogie, il convient pour mener à bien des études de "[…] *gather evidence about the effects of instructional conditions of instruction*" (Chapelle, 2004 : 594). Cela implique dans un premier temps de rassembler les éléments qui ont caractérisé le dispositif pédagogique, dont le scénario lui-même. Or, nombre de publications scientifiques rendent compte de façon incomplète de ces éléments. Des traitements sont ainsi appliqués sur des données d'apprentissage, en quelque sorte décontextualisées, et conduisent les auteurs à tirer des conclusions sur certains dispositifs de formation ou environnements technologiques, conclusions que d'autres travaux viendront ensuite contredire (parce que, par exemple, de meilleurs scénarios pédagogiques auront été utilisés). **Du point de vue méthodologique,** il convient de relier ces différents types de données pour avoir un objet digne d'analyse, comme le souligne cet extrait à propos des interactions produites dans les forums de discussion :

> *La recherche sur le forum de discussion en contexte éducatif tente de rendre compte de phénomènes complexes à l'aide de méthodes d'analyse de contenu qui n'éclairent qu'un aspect de la réalité. Pudelko, Daele et Henri (à paraître) en résument les limites ainsi : « la principale difficulté de ce type d'analyse de contenu provient de son objectif même : rendre compte de la dynamique interactionnelle dans la conversation médiée par ordinateur dans ses relations avec les processus sociocognitifs mis en jeu par les interactants. Pour y arriver, la méthode d'analyse devrait être capable de traiter le discours comme une interaction verbale située, dans ses dimensions linguistiques (liées à la syntaxe, au lexique et à la sémantique), situationnelles (liées à l'univers de référence et à la situation d'interaction) et des contraintes hiérarchiques (liées à la structure hiérarchique du discours). (Henri & Charlier, 2005)*

Les différentes dimensions de ces interactions situées (linguistiques, situationnelles et liées aux contraintes imposées par le medium de communication) nécessitent que les données recueillies dans les expérimentations le soient selon un protocole assurant l'exhaustivité, et qu'elles soient organisées de façon à permettre :

- de situer la lecture des traces dans le contexte de la situation d'apprentissage
- de saisir les contraintes de l'environnement les ayant générées
- d'appliquer des analyses automatisables sur des données numériques standardisées. Ces conditions permettent d'envisager le développement d'outils de suivi de la formation par les acteurs (tuteur ou apprenants) tels que des tableaux de bord.

Symposium EPAL

L’objectif du projet Mulce est de construire un modèle conceptuel permettant d’organiser un corpus en éléments indexés, reliés sémantiquement et dans un format suffisamment ouvert pour permettre à différents outils de lire ces données pour en faire leur analyse qualitative ou quantitative.

Une partie de ce modèle, pour ce qui est de la caractérisation du dispositif de formation devra s'inspirer des nombreux travaux en cours des communautés qui visent à étendre l'usage des technologies d'aide à l'apprentissage en insistant en particulier sur l'adoption de **modèles décrivant les scénarios pédagogiques** comme IMS-LD (*Instructional Management Systems-Learning Design*) (IMS, 2003 ; Lejeune & Marino, 2005). Un pont serait ainsi fait pour la première fois au niveau de la recherche, entre ce qui était planifié pédagogiquement et les traces de ce qui s'est déroulé, traces au milieu desquelles on peut espérer repérer les effets en terme d'apprentissage. Mais les modèles du type IMS-LD ont besoin d'être adaptés car, comme nous l'avons signalé dans notre définition, c'est le descriptif du dispositif de formation tel qu'il s'est réellement déroulé et non simplement planifié qui a besoin d'être intégré au modèle du corpus d'apprentissage.

Les dispositifs de recherche ou la dynamique du dispositif de formation

Disposer de traces des interactions est une chose, encore faut-il pouvoir les replacer dans la dynamique de la situation d'apprentissage. Des dispositifs de capture multimédias, dont des vidéogrammes, sont alors des ressources de première importance pour les chercheurs, ainsi qu'en témoigne en apprentissage des langues ce numéro thématique de la revue System (2004) "*Incorporating multimedia capability in the reporting of applied linguistics research*". Ces données capturent à travers la vidéo d'un écran la simultanéité des opérations provenant d'un seul participant agissant en différents endroits d'un système ou de celles provenant des autres participants. Elles peuvent aussi, suivant les besoins, permettre de saisir ce qui se passe dans l'environnement en ligne ou matériel des participants (Smith& Gorsuch, 2004, par exemple). Dans un désir de circonscrire notre espace de projet, on notera également l'existence de recherches sur les interactions d'utilisateurs agissant autour d'environnements technologiques. La vidéo vient alors capturer leur gestuel (notamment) devant un système. On est alors plus orienté vers des problématiques relevant de l'ergonomie, de l'utilisabilité (*usability*) des systèmes (voir les travaux de Mondada et son équipe (Crego, 2004)).

Dans le projet Mulce, même si nous n'écartons pas l'intérêt des vidéogrammes capturant l'environnement matériel, nous nous intéressons prioritairement aux vidéogrammes de capture des écrans car ils contiennent non seulement la mise en scène complète des interactions survenant en ligne (seul espace partagé par les acteurs en situation) mais les données même des interactions (dialogues oraux, mouvements, actions de production collaborative, etc.), dont de larges composants multimodaux doivent être extraits à la main (suivant une méthodologie qui est l'un des objectifs de ce projet de recherche) car ils n'existent pas en tant que traces dans des fichiers comme c’est au contraire le cas du clavardage (*chat*) par exemple.

Multimodalité et corpus d'apprentissage

La multimodalité est présente à plusieurs niveaux dans un corpus d'apprentissage : (pour une définition des notions de "média", "mode", "modalité" et "multimodalité" dans nos contextes, on se reportera à (Chanier & Vetter, 2006)) :

- Les données résultats d'une formation (Da) peuvent être de nature textuelle, orale/audio, graphique, iconique, etc. Cette diversité provient des différents modes de communication et de production (écrit, audio, dessin, vote, etc.) que des environnements informatiques intégrant la multimodalité, mettent à disposition des apprenants pour s'exprimer (Reffay & Betbeder, 2006).

Symposium EPAL

- Les données provenant du dispositif de recherche (Dc) peuvent se présenter sous forme de vidéogrammes dont il faudra ensuite extraire les différentes modalités présentes en particulier dans les environnements synchrones.

Les premières étapes de la recherche consisteront à annoter et transcrire ces données avant de se livrer à des analyses multiples (pour avoir une idée de cette variété, voir par exemple (Herring, 2004) et sa recension des perspectives d'analyses de la notion de communauté en ligne). C'est pourquoi le projet Mulce s'intéresse à l'élaboration à la fois de modèles permettant d'intégrer le niveau primaire et les différents niveaux de description (Salmon-Alt, Romary & Pierrel, 2004) et également d'outils, modèles et procédures aidant à produire ces descriptions (Kress & Jewitt, 2001 ; Avouris & al., 2004 ; Levine & Scollon, 2004 ; Baldry & Tibault, 2006 par exemple).

Garantir l'accès libre et l'interopérabilité des corpus

L'accès ouvert aux résultats de la recherche, que ce soit les publications (Chanier, 2004) ou les données, est un enjeu actuel de première importance, comme le rappelle cet extrait du Conseil de Recherche en sciences humaines du Canada :

> *Facilitate the advancement of knowledge in the social sciences and humanities by encouraging researchers to share research data. Sharing data strengthens our collective capacity to meet academic standards of openness by providing opportunities to further analyze, replicate, verify and refine research findings. […] Finally, researchers whose work is publicly funded have a special obligation to openness and accountability […] Research data includes quantitative social, political and economic data sets; qualitative information in digital format; experimental research data; still and moving image and sound data bases; and other digital objects used for analytical purposes. SSHRC (2005)*

Dans notre domaine de recherche, la non accessibilité aux données de recherche, qui est l'état de fait quasi général au sein de notre communauté internationale, est un frein de premier ordre à la reconnaissance des situations d'apprentissage en ligne comme un objet d'étude scientifique : elle empêche les vérifications ou infirmations, la réplication, le raffinement, les analyses multiples etc. Les cas de réanalyse de données d'apprentissage sont tellement rares qu'on n'hésitera pas à citer l'étude de Kramsch et Thorne qui à partir des données de Kern ont produit une interprétation différente du premier pour expliquer l'échec en terme de compétence communicative dans les échanges entre apprenants de L1 différentes (Kern, Ware & Warshauer, 2004 : 251).

Plusieurs types de facteurs expliquent cet état de non accessibilité. Tout d'abord des habitudes de travail où l'on partage les discours sur les produits de la recherche sans interroger les démarches méthodologiques, où l'on pratique l'autarcie et la rétention d'information à des fins mal comprises de reconnaissance scientifique. Mais pour que des données de recherche puissent être rendues accessibles, il faut aussi que :

- Ces données soient **lisibles et pérennes** et donc extraites des environnements propriétaires au sein desquels elles résident (sinon elles sont dispersées, incomplètes, interdites aux chercheurs ne disposant pas d'accès dans ces environnements ; à durée de vie limitée à celle de la version du logiciel, etc.). Le travail d'extraction peut être direct dans certains (rares) environnements ou nécessiter des interventions techniques notables (voir le cas de notre corpus *Simuligne* et de la plate-forme WebCT (Reffay, 2001)).
- Qu'elles soient **homogènes** ce qui suppose des pré-traitements. Les données brutes sont peu souvent utilisables directement. Il faut souvent opérer des montages entre données audio et vidéo, changer des formats de fichiers, réorganiser les noms et structures des fichiers et répertoires, les cataloguer pour en faciliter l'accès, etc.

Symposium EPAL

- Qu'elles soient **valides**. La validité s'estime ici, d'une part, par le caractère plus ou moins exhaustif des données recueillies (le scénario pédagogique et le dispositif de formation doit garantir que la très grande majorité des interactions ne s'est pas déroulée en dehors du système sinon les conclusions du travail de recherche pourront difficilement être étayées) et, d'autre part, par le caractère écologique de la situation d'apprentissage (qui s'apprécie au regard de la durée de la formation, du nombre de sujets, de la qualité de la préparation, des activités, etc. toutes choses éloignées des conditions expérimentales artificielles rencontrées dans les laboratoires).

Les méthodes de travail en usage satisfont peu fréquemment ces conditions, ce qui explique la rareté actuelle de collections de données qui pourraient devenir des corpus d'apprentissage. Pour se lancer dans le projet Mulce, il fallait donc que les équipes partenaires disposent au préalable de telles collections (cf. infra avec *Simuligne*, *Copéas*, *Tridem*).

L'**accès libre et l'échange** de corpus d'apprentissage au sein de la communauté internationale sera alors garanti par :

- La transformation de la collection des données en véritable corpus d'apprentissage, par adjonction, notamment, des données décrivant le dispositif et donc élicitant le contexte pour les équipes de recherche n'ayant pas vécu la situation de formation.
- L'organisation du corpus suivant un modèle explicatif des points précédents et utilisant des langages et formats de description compatibles avec les standards, voire les normes. Au niveau des données, cela suppose qu'elles utilisent largement l'encodage XML avec un respect d'un côté des principes d'organisation des corpus linguistiques en ligne (cf. (Salmon Alt, Romary & Pierrel, 2004) et standards déjà cités, TEI Iso/TC 37/SC4) et, de l'autre, de ceux orientés vers la conception pédagogique (IMS-LD, déjà cité).
- L'interopérabilité des formats et systèmes d'accès. Pour que ces corpus soient utilisables sur les différents systèmes informatiques, ils doivent utiliser des formats non propriétaires et multi-systèmes d'exploitation. L'accès ouvert sera garanti par l'utilisation de protocoles de communication adéquats, notamment celui des archives ouvertes OAI-MPH (2002) et de métadonnées compatibles avec le Dublin Core et sans doute aussi les recommandations de l'*Open Language Archives Community* (Olac, 2006).

Le serveur Mulce devra satisfaire ces conditions techniques pour offrir les garanties de visibilité scientifique.

Même si cela nous semble aller de soi, nous insistons sur le terme "accès libre" qui doit bien être interprété ici comme synonyme non seulement d'un accès ouvert à tous les systèmes (cf. interopérabilité), mais également d'un accès immédiatement libre et gratuit, ainsi qu'offrant à terme des garanties d'archivage et de conservation. Il s'inscrit dans l'esprit des archives ouvertes (*open archives*) et dans le respect de la notion de "**contribution scientifique ouverte**" telle qu'exigée par nos directions de recherche (voir notamment Berlin, 2003).

Questions éthiques et contrat de cession des droits

La double contrainte d'accès libre et de rassemblement de données impliquant des acteurs en situation de formation pose un défi, comme le rappelle cet extrait :

> *Any discussion of technology in second language research would not be complete without raising the ethical challenges that researchers face in SLA [Second Language Acquisition] research in general and particularly in research involving the collection and archiving of personal performance data that reveal personal attributes (Chapelle, 2004 : 599).*

Symposium EPAL

Pour relever ce défi, le projet Mulce doit trouver les procédures qui offrent un compromis entre les garanties d'anonymat des personnes, leur automatisation, la cohérence des données après traitement. La question de la cession des droits par les participants de leur production aux fins de diffusion doit être abordée, ainsi que celle des droits associés aux ressources pédagogiques incluses dans le corpus.

L'élaboration d'un contrat associé aux corpus déposés et aux utilisations fera l'objet d'un travail conjoint avec nos partenaires internationaux. Mais signalons déjà, que nous disposons de contrats de cessions des droits signés par tous les participants concernés par nos trois corpus et que *Tridem*, particulièrement, est un bon cadre de discussions puisque 3 institutions de 3 trois pays différents y sont associées, chacune disposant de son propre contrat d'éthique pour ses apprenants.

2.2 Les collections de données à notre disposition

Nous donnons des informations sur les 3 collections de données déjà en notre possession en indiquant les projets de recherche et les dispositifs de formation qui ont présidé à leur naissance. Dans le cadre du projet Mulce ces collections de données devront bien sûr être transformées en corpus d'apprentissage, au sens défini précédemment. Cette tâche figure dans notre agenda.

Collections de données Simuligne (asynchrone – 2001)

Soutenu par le ministère français de la recherche dans le cadre du programme Cognitique, le projet ICOGAD (Interactions COgnitives en Groupe et A Distance, 2000-2002) a monté en partenariat avec le département de langues de l'Open University du Royaume-Uni, une expérimentation en formation à distance nommée « *Simuligne* » basée sur la technique de la « simulation globale » habituellement utilisée en classe présentielle intensive de langue. Cette formation a concerné 40 apprenants anglophones de l'Open University, 4 tuteurs, 10 étudiants natifs (francophones) servant de soutien et de référents pour la langue et la culture française. Ayant pour objectif l'apprentissage de la langue française pour des personnes d'un niveau intermédiaire à avancé, elle s'est déroulée entièrement à distance, sur une durée de 10 semaines et l'ensemble des ressources et médias de communication étaient concentrés sur une seule plate-forme de téléformation asynchrone (WebCT).

Collections de données Copéas (synchrone – début 2005)

Le projet de recherche *Copéas* (COmmunication Pédagogique en Environnement orienté Audio Synchrone) concerne une expérimentation écologique en vue de l'analyse des corpus d'interactions multimodales (audio, graphiques, textuelles). *Copéas* est une sous-partie du projet ODIL, ACI "Éducation et Formation" 2004**.** La formation s'est déroulée sur 8 semaines début 2005, dans l'environnement audio-graphique synchrone Lyceum, développé à l'Open University (R-U). Elle vise à développer des compétences d'expression orale dans un contexte professionnel en anglais langue seconde chez 14 apprenants de niveau linguistique hétérogène, inscrits en master professionnel FOAD (Université de Franche-Comté).

Collections de données Tridem (asynchrone et synchrone – fin 2005)

Trois partenaires institutionnels : Carnegie Mellon University (EU), The Open University (GB) et l'Université de Franche-Comté ont monté une formation en langue à distance. Celle-ci a eu lieu fin 2005, sur 10 semaines, et mêlait deux dispositifs d'apprentissage collaboratif : l'un asynchrone et essentiellement écrit et graphique groupant les apprenants par *Tridem* anglo-américano-français : le blog, l'autre multimodal et synchrone et rassemblant des petits groupes de 7 ou 8 personnes dans l'environnement Lyceum.

Synthèse Epal

Données sur les 3 corpus

	Simuligne	Copéas	Tridem
Participants	- 1 coordinateur (UFC), - 10 natifs (UFC), - 40 apprenants (OU) - 4 tuteurs (OU) - 4 groupes de 12 en parallèle - un groupe les regroupant tous pour les activités interculturelles.	- 14 apprenants (UFC) - 2 tuteurs (OU) - 2 groupes	- 28 apprenants : -- 13 étatsuniens (CMU), -- 10 français (UFC), -- 5 britanniques(OU) - 10 tridems, - 5 tuteurs (OU, CMU, UFC)
Environnements technologiques	- Plate-forme asynchrone (WebCT) - Formulaires sur serveurs (Php + Mysql).	- Plate-forme Audio-graphique synchrone : (Lyceum) - Plate-forme asynchrone (WebCT)	- Blog - Plate-forme Audio-graphique synchrone : (Lyceum)
Da) Les données résultats d'une formation			
Interactions Devoirs rendus	- 2686 messages forum, - 4062 courriels - 5680 tours de clavardage - 93 documents textuels, - une image - 28 fichiers audio	- 5506 tours de parole audio (qui représente en temps cumulé 8h29) - 1529 tours de clavardage - interactions multimodales à transcrire sur les 16 séances Lyceum	-11 blogs archivés (avec 610 messages et 127 photos), - 19 séances audio-synchrones montées - 1030 tours de clavardages - 8 évaluations individuelles
Productions affichées	342 pages web incluant 115 images et 44 fichiers audio	Documents, cartes conceptuelles et tableaux blancs non comptés	- 10 documents, - 4 cartes conceptuelles - 51 tableaux blancs issus des séances Lyceum
Db) Les données extraites du dispositif de formation			
Ressources pédagogiques	569 fichiers représentant 75 pages web (en plus des consignes)	- un guide spécifique aux tuteurs, - un guide spécifique aux apprenants	un guide spécifique aux apprenants
Scénario	28 activités réparties en 7 étapes sur 12 semaines, présentées par: - 41 fiches de consignes aux apprenants, - 21 fiches spécifiques pour les tuteurs - 14 spécifiques aux natifs	8 activités sur 10 semaines	4 activités sur 10 semaines
Dc) Les données provenant du dispositif de recherche			
Questionnaires, entretiens	12	- 14 questionnaires apprenants, - entretiens semi dirigés (x apprenants, 1 tuteur), - 9 Critical event recall (8 apprenants, 1 tuteur)	- 26 pré-questionnaires, - 13 post-questionnaires, - 13 post-entretiens dirigés (1 apprenant, 1 tuteur)
Taille	Total : 650 Mo : - 30 000 fichiers répartis dans 2708 dossiers	Total : 35,3 Go : - 37 vidéos (27h) - 512 autres fichiers répartis dans 117 dossiers.	Total : 7,37 Go : - 16 vidéos (20h) - 939 fichiers
Cession des droits	Oui	Oui	oui

Symposium - Eros

2.3 Notre expérience en FOAD, en transcription et analyses des interactions

Dans le cadre de recherches sur les dispositifs d'apprentissage pour la formation à distance, nous nous intéressons à l'étude des interactions à l'intérieur d'une communauté formée d'apprenants, de formateurs et de tuteurs. Une attention particulière est apportée pour que ces études gardent un ancrage sur le terrain, elles sont effectivement supportées par des

données issues de réelles situations d'apprentissage. Dès 1999, l'équipe bi-disciplinaire de Besançon s'inscrit dans ce domaine de recherche en se positionnant de la façon suivante :

A l'opposé d'autres travaux en SHS, nous ne cherchions pas à comparer les apprentissages en ligne versus ceux en présentiel, mais plutôt à explorer les types d'interactions permises par les TIC dans des dispositifs d'apprentissage à construire (*Simuligne*, *Copéas*, *Tridem*).

Du côté de l'informatique, l'équipe ne cherchait pas à construire une nouvelle plate-forme, mais plutôt à en adopter une qui soit suffisamment répandue et utilisée, pour ne pas reconstruire les services de base déjà éprouvés dans les plates-formes commerciales, mais plutôt proposer de nouveaux outils/services visant à suivre, analyser ou favoriser les interactions.

Le Creet, étant à l'intérieur même de l'Open University (Royaume-Uni) est directement au contact de nombreuses formations à distance. Le Laseldi dirige un master professionnel FOAD (2006) et certains membres du LIFC effectuent une part de leur enseignement à distance et en ligne. Le fait de partager cette proximité au terrain de la FOAD doit être vu comme un garant de l'intérêt et de l'utilité de ces nombreux travaux. Ils portent en particulier sur l'incidence des environnements technologiques et des scénarios pédagogiques sur la nature et la qualité des interactions et donc de l'apprentissage (Lamy, 2006 ; Lamy, 2004) le développement de dispositifs et de ressources éducatives favorisant le travail collaboratif (Reffay & Chanier, 2001 ; Betbeder & Tchounikine, 2005), les rôles tenus par le tuteur (Greffier, 2005), les stratégies d'apprentissage (Jeannot et al. 2006), l'évaluation d'un dispositif et l'analyse des usages (Reffay & Betbeder, 2006 ; Chanier & al., 2006 ; Lamy & Hassan, 2003), les indicateurs structurels ou quantitatifs de suivi d'un groupe collaboratif d'apprentissage à distance et en ligne (Greffier & Reffay, 2006) et en particulier ceux en rapport avec l'analyse des réseaux sociaux (Reffay & Chanier, 2003 ; Reffay & Lancieri, 2006), les outils de transcription d'actions multimodales dans un environnement synchrone (Betbeder & al., 2006), les outils d'analyse et/ou de suivi des interactions (Mbala & al., 2002 ; Mbala& al., 2003) et enfin l'émergence des besoins de spécification de corpus d'apprentissage pour leur partage (Noras, 2006), point de départ de ce projet.

On voit ici que de nombreux aspects des situations d'apprentissage collaboratif (d'abord asynchrone : *Simuligne*, puis synchrone : *Copéas*, puis mixte : *Tridem*) sont analysés dans ces différents travaux. Les différentes modalités temporelles (asynchrone et synchrone) impliquent des protocoles de recueil des données très différents. Par nature, une plate-forme synchrone conserve l'essentiel des informations d'une visite à l'autre pour permettre aux divers acteurs de partager les productions ou les interactions. Les données sont donc nécessairement conservées, mais pas toujours lisibles (format propriétaire), après arrêt ou changement de la version de la plate-forme utilisée. Des programmes de transformation sont donc nécessaires pour pouvoir lire ces données a posteriori et indépendamment de la plate-forme, condition nécessaire pour assurer leur pérennité. A l'opposé, les plates-formes synchrones ont pour priorité la fluidité des flux audio ou vidéo, ce qui nécessite en général une grande puissance de calcul et une large bande passante. La sauvegarde des informations devient secondaire et bien que certaines facilités soient proposées par les plates-formes telles que Lyceum ou Centra, les traces qui y sont proposées ne contiennent pas l'intégralité des informations utiles à la description minutieuse des interactions pour un chercheur. C'est pourquoi nous avons eu recours aux enregistrements des vidéos pour *Copéas* et *Tridem*. Les protocoles de recueil des données nous ont permis d'améliorer l'exhaustivité d'une collection de données. L'échange de ces collections de données entre partenaires nous a montré combien il est difficile et important de décrire chaque partie pour la rendre utile et accessible. Nos travaux sur la mesure de la cohésion des groupes à partir de l'analyse des réseaux sociaux (dans *Simuligne*) nous ont tout particulièrement amenés à traiter les données de manière automatique, ce qui nécessite un haut degré d'homogénéité dans la collection des données. Nos travaux plus récents sur la transcription d'actions multimodales, passant par des phases manuelles, ont mis en évidence certaines

Symposium EPAL

incohérences dans les transcriptions. Nous avons donc entamé une série de tests de cohérence de la base de données afin de les détecter et de les réparer systématiquement.

La mutualisation des expérimentations entre le LIFC, le Laseldi et le Creet a déjà permis d'effectuer des analyses de natures différentes sur les mêmes collections de données (*Simuligne*, *Copéas*). Il est temps maintenant que ces collections de données, issues de dispositifs novateurs et donnant lieu à de nombreuses et riches interactions, puissent être analysées par un plus large ensemble de chercheurs pour permettre en particulier à d'autres méthodes et d'autres outils de confronter leurs résultats aux nôtres en vue de faire progresser notre compréhension négociée des phénomènes d'apprentissage collaboratifs en ligne. Ce premier pas vers la mutualisation de corpus d'apprentissage est selon nous une étape décisive pour la maturation scientifique de notre domaine (CALL, SLA, EIAH, ITS, CSCL) et nous ferons notre possible pour qu'il reçoive l'adhésion de nos collaborateurs extérieurs afin qu'ils déposent de nouvelles analyses pour enrichir nos corpus ou même de nouveaux corpus d'apprentissage.

2.4 Outils d'aide à la transcription et à l'analyse

Deux partenaires du projet, en l'occurrence le LIFC et le LIP6 ont une expérience significative en développement d'outils d'annotation / transcription et en outils d'analyse. Nous la rapportons ici en ouvrant des perspectives pour le projet Mulce.

2.4.1 LIFC : Outils d'annotation/transcription et analyse de patterns

En juin 2005 s'est tenu un atelier de travail sur la comparaison d'outils d'annotation multimodaux dans le cadre du *Second Congress of the International Society for Gesture Studies* (ISGS, 2005). Six outils : ANVIL, ELAN, EXMARaLDA, Media & Text Editors, TASX et MacVisSTA y ont été comparés sur quatre corpus donnés sous forme de vidéogrammes aux contenus variés (conversation libre, narration d'histoire, description d'un itinéraire routier, travail collaboratif d'organisation) (Rohlfing & al., 2005).

D'un côté, les outils issus de SHS cités ci-dessus supposent des actions spécifiques et produisent le plus souvent des représentations sous forme de partitions, qui permettent une lecture très précise de certains passages critiques, mais ne sauraient servir l'exhaustivité en vue de réaliser des analyses quantitatives. De l'autre, l'outil ColAt (Avouris et al., 2004), basé sur la théorie de l'activité sert plutôt à l'annotation et à l'analyse qu'à la transcription, puisque l'équipe HCI ayant la main sur les outils de collaboration (Synergo en particulier), peut générer les traces automatiquement. C'est pourquoi nous avons choisi de développer un prototype de transcription d'actions multimodales en environnement synchrone : *Tasync*.

Les vidéos enregistrées (dans *Copéas* et *Tridem*) correspondent à un film de l'écran partagé au cours de la séance, i.e. on voit l'auteur, le lieu, l'heure et le type de toutes les actions des différents acteurs. Cependant ce format est un « fondu à plat » des actions réalisées au sens ou aucune de ces actions n'est indexée. Ceci interdit tout traitement informatique pour rechercher ou comptabiliser des occurrences, repérer des schèmes d'actions, etc. Nous avons donc développé l'outil *Tasync* (Djouad, 2005 ; Betbeder & al., 2006) pour aider la transcription des actions à partir de vidéogrammes d'expérimentation. L'élément central de *Tasync* est le modèle des différentes actions multimodales réifié dans la structure de la base de données. Injecté dans cette structure par la procédure de transcription, le vidéogramme initial se transforme en une base de données typée des actions interrogeables.

Enfin on notera cet autre travail du LIFC dans le domaine de l'analyse. Dans le cadre d'un stage de master en informatique, est développé actuellement un service (*PADIS* : outil de "*pattern discovering"*) (Betbeder & al., 2007) à intégrer au serveur Mulce, et/ou à diffuser comme outil. *PADIS* a pour objet de découvrir automatiquement les séquences d'actions

Symposium Final

récurrentes pour soumettre d'éventuels schèmes à l'analyse, en particulier, des chercheurs de SHS.

2.4.2 LIP6 : Des outils d'analyse textuelle

Le LIP6 a mis en place des outils d'analyse textuelle qui s'avèrent pertinents, appliqués à des domaines de recherche en Sciences humaines tels que la critique génétique textuelle ou la didactique du Français langue étrangère. L'analyse textuelle et le Traitement Automatique des Langues peuvent tirer partie de nouvelles techniques d'apprentissage symbolique et de fouille de données. Conçu au LIP6 par Jean-Gabriel Ganascia, le *Littératron* (Ganascia 2001) extrait automatiquement des motifs syntaxiques récurrents à partir de textes écrits en langage naturel ; associé à un analyseur syntaxique de textes en arbres, il révèle les singularités stylistiques d'un auteur ou d'un genre. Utilisé en Sciences du langage, dans le domaine de l'acquisition en langue étrangère du français écrit, le *Littératron* est capable d'effectuer un diagnostic linguistique de l'apprenant, en révélant certaines compétences morpho-syntaxiques présentes ou absentes (Audras & Ganascia, 2006). D'autre part, il nourrit le questionnement autour de la notion d'évaluation du style en langue étrangère en proposant une mesure objective de la qualité stylistique d'une production d'apprenant.

Medite (Machine pour l'Etude Diachronique et Interprétative du Travail de l'Ecrivain) est un aligneur textuel, mis au point par Jean-Gabriel Ganascia au LIP6 (Ganascia & Bourdaillet, 2006), qui recense systématiquement toutes les transformations qui font passer d'un état textuel à un autre (déplacements, suppressions, insertions, remplacements). Conçu pour l'aide à l'analyse de versions de manuscrits d'écrivains en génétique textuelle, cet outil est d'ores et déjà largement utilisé dans ce domaine. Dans le domaine de l'acquisition des langues étrangères, cet outil semble utile pour la comparaison de versions de productions d'apprenants. En effet, il révèle les invariants et les suppressions, insertions et/ou déplacements d'expressions d'une version à une autre, qui sont autant de traces laissées par l'apprenant sur sa démarche d'apprentissage. De plus, l'analyse automatique des motifs syntaxiques les plus fréquemment déplacés, remplacés et/ou supprimés, outil couplé à *Medite*, renseigne l'apprenant et/ou l'enseignant sur la qualité stylistique de la production ainsi que sur les points de morpho-syntaxe dont l'emploi est encore hésitant.

SEEK (Système Expert d'Exploration (K)Contextuelle), conçu par C. Jouis (Jouis, 1995) au laboratoire LaLICC (Langages, Logiques, Informatique, Cognition et Communication, UMR 8139, Université Paris-IV Sorbonne – CNRS) est un outil d'extraction de relations sémantiques entre termes dans les textes. Cet outil a permis la modélisation des connaissances par les cogniticiens à partir d'entrevues d'experts dans des domaines variés. Appliqué à ce projet, il pourra servir, par exemple, à renseigner sur les champs sémantiques présents dans les productions d'apprenants.

Le LIP6 propose dans Mulce de mettre à disposition les outils d'analyse textuelle présentés et d'en développer de nouveaux, notamment dans le sens d'une analyse sémantique de la production. La compétence de l'équipe ACASA en Intelligence Artificielle mise à disposition pour l'extraction automatique d'informations sur un texte, ou plusieurs textes, complète celle des autres équipes partenaires. Les outils automatiques conçus au LIP6 apportent une analyse du corpus d'apprenants qui est capable d'évaluer la production en termes de compétences linguistiques et qui propose, par conséquent, d'autres critères d'information sur les corpus recueillis.

2.5 Agenda

Pur faciliter la collaboration des membres du projet Mulce avec des partenaires extérieurs, nous présentons les objectifs dans l'agenda du projet Mulce, année par année.

Symposium EPAL

Année 1 : janvier 2007 – décembre 2007

Objectifs : conception, développement et mise en place du niveau de description primaire d'au moins deux échantillons de corpus (l'un synchrone, l'autre asynchrone) et validation de ce niveau de description.

Année 2 : janvier 2008 – décembre 2008

Objectifs : intégration dans le modèle de description primaire des données vidéo et dans les modèles de transcription, des interactions multimodales. Développement et adaptation des outils de d'analyse textuelle et de recherche plein texte dans les corpus.

Année 3 : janvier 2009 – décembre 2009

Objectif : premiers retours de collaborations (analyse de corpus Mulce ou dépôt de nouveaux corpus dans mulce) et intégration pour diffusion.

3 Premières avancées et coopérations

3.1 Structuration d'un corpus d'apprentissage dans Mulce

Engagés depuis plus d'un an dans cette réflexion, nous avons déjà partiellement dessiné les contours de la structure Mulce. D'un point de vue conceptuel, nous avons déjà défini plusieurs parties nécessaires à la composition d'un corpus de formation en ligne (Noras & al., 2007, Reffay & al. 2007) :

- Les traces et transcriptions : cœur du corpus ;
- Leur contexte : scénario pédagogique et protocole de recherche ;
- Les licences : contrats de cession des droits et de protection des acteurs et contrat d'utilisation à des fins de recherche ou d'enseignement.

Symposium Epal

Avant de stabiliser une spécification qui soit suffisamment simple pour être réutilisable, riche pour être expressive et donc utile aux analyses, nous avons d'abord étudié comment nous pourrions réutiliser certaines spécifications existantes dans le monde de la formation à distance et en particulier IMS-CP et IMS-LD. Nous avons donc repris le scénario de Simuligne et l'avons entièrement décrit dans le logiciel Reload (éditeur LD) en utilisant ainsi la spécification IMS-LD. Nous avons ainsi été confrontés aux difficultés et limites de la métaphore théâtrale proposée par cette spécification. Cette description détaillée au niveau conceptuel du scénario, définissant la formation dans sa structuration, son séquencement en différentes étapes et activités en précisant systématiquement les rôles, lieux, documents services et outils, nous permettra de lier les interactions et productions recueillies pour rendre le contexte de l'activité, telle qu'elle a été conçue et présentée aux acteurs. Cette association au contexte était, dès l'origine du projet Mulce, considérée comme une nécessité puisque les données sont susceptibles d'être interprétées par des chercheurs n'ayant ni conçu ni vécu la situation d'apprentissage. Une simple chronologie des traces n'est pas à même de renseigner sur le contexte de la tâche en cours. De ce point de vue, nous considérons que IMS-LD propose une structure suffisamment riche et détaillée pour contenir toute l'information utile à la description du scénario. En revanche, le résultat de cette description, lorsqu'il s'agit d'une formation aussi complexe que Simuligne, est assez difficile à lire et à appréhender quand on ne connaît pas précisément IMS-LD. Il est cependant possible de simuler le déroulement du scénario dans le logiciel Reload Player. Nous pensons que des formes plus visuelles et plus légères (telles que celle proposée par Mot+) seraient plus adaptées à la complexité de telles formations pour être appréhendées par des non spécialistes.

3.2 Le défi de l'anonymisation

D'un point de vue pratique, dès lors que les données sont susceptibles d'être accessibles, nous devons garantir une certaine protection des acteurs de la situation et en particulier rendre les données (et transcriptions) anonymes. Cette **anonymisation** doit cependant conserver la cohérence et la lisibilité des interactions. L'identification des acteurs (auteur d'un message de forum, de courriel ou d'un tour de parole de chat) fournie par la plateforme de téléformation étant systématique, nous pouvons la remplacer par un code ou un pseudonyme de manière tout aussi systématique. En revanche, dans le contenu des interactions, le nom des acteurs est souvent présent dans une interpellation (d'autres acteurs que l'auteur) ou une signature (le nom de l'auteur lui-même). Naturellement, dans ce cas, la syntaxe utilisée par les acteurs pour se nommer est tout sauf systématique : ils peuvent utiliser des diminutifs, des surnoms, des abréviations, du nom ou du prénom et, par le jeu des erreurs de typographie, fabriquer de nombreuses formes syntaxiques pour représenter une même personne. L'enjeu est donc de repérer chacune des formes, de la caractériser afin de ne pas perdre d'information et de produire une forme structurée de remplacement qui empêche l'identification de l'acteur réel tout en conservant le potentiel d'intention de la forme choisie (diminutif, surnom, etc.), ou le potentiel d'ambiguïté porté par cette forme, e.g. : si l'auteur parle de « Chris » alors qu'il y a parmi les acteurs « Christophe » et « Christian », il se peut que le choix de cette forme induise des quiproquos qui ne seraient plus compréhensibles si nous remplacions (arbitrairement) « Chris » par le code correspondant à l'acteur « Christian ». Le défi qui s'offre à nous est donc de définir la structure, les outils et les procédures de traitement automatique qui permettent de conserver l'information dans toutes ses nuances sans dévoiler l'identité réelle de l'auteur.

Un prototype permettant de remplacer automatiquement les formes syntaxiques repérées identifiant les acteurs réels par des formes équivalentes mais fictives a été développé au LIUM (Laboratoire d'Informatique de l'Université du Maine, au Mans). Il est issu d'une collaboration avec le LIFC, qui a prêté son corpus et d'une réflexion conjointe qui sera présentée comme poster et démo à la conférence EIAH'2007, à Lausanne à la fin du mois de juin (Reffay & Teutsch, 2007).

Symposium Epal

3.3 Comment peut-on coopérer ?

Le symposium « Corpus » du colloque EPAL est l'occasion de présenter à la communauté scientifique du domaine de l'apprentissage en ligne, notre manière d'aborder la question du partage des corpus. Nous attendons également de ce rendez-vous, un retour de la communauté sur l'intérêt qu'elle porte aux objectifs de Mulce, mais aussi son point de vue sur les choix que nous faisons sur la manière de structurer les corpus, et, par conséquent, le type de traitements et d'analyses qui peuvent être envisagés sur les corpus mis à disposition par le projet Mulce. Nous espérons de cette rencontre qu'elle recense les types d'analyses que les chercheurs de la communauté sont déjà en mesure de faire sur leur propre corpus et qu'elle apporte une réflexion sur le chemin à parcourir pour que ces analyses soient applicables à d'autres corpus. Enfin, si de nombreux outils de traitement ou d'analyse doivent pouvoir s'appliquer, à terme, à chacun des corpus de Mulce de façon individuelle, nous souhaitons également évoquer de nouvelles pistes de recherche qui s'ouvriront par l'intégration de ces différents corpus dans une même base.

C'est donc bien au moment où le projet Mulce définit les choix qui contraindront l'intégration des nouveaux corpus que nous avons besoin du retour de la communauté pour corriger, adapter ou enrichir nos spécifications, pour les rendre d'une part suffisamment souples et adaptables au plus grand nombre de corpus issus de situations d'apprentissage collaboratif en ligne, et d'autre part, suffisamment formelles pour être opérables par les méthodes et outils d'analyse ou de traitement automatique existants ou à venir.

Le projet de recherche Mulce a déjà fixé ses objectifs, ses membres et son agenda sur trois années. Mais ce projet suppose l'adhésion de nouveaux partenaires dans un cadre plus

large pour que cette impulsion devienne véritablement utile à la communauté. Le véritable but de Mulce ne se limite pas à la spécification d'une structure de corpus accessible par d'autres (même si cela est déjà difficile), mais à la définition d'un socle de partage qui permette à la communauté scientifique d'aller plus loin dans la comparaison, la validation et la réutilisation d'analyses sur des données effectivement partagées. Si le projet Mulce se donne trois ans pour mettre en place une première version de ce socle, il appartient à chacune des équipes de la communauté de monter de nouveaux projets qui lui permettent de profiter pleinement de ce socle en partageant ses corpus, ou en construisant de nouvelles analyses concurrentes, complémentaires ou comparatives sur les corpus déjà mis à disposition par ce socle.

4 Bibliographie et état de la question

Les liens Internet indiqués dans cette section de références (bibliographie, sites, etc.) ont été vérifiés en date du 2 juin 2007.

Bibliographie

Audras, I. & Ganascia, J-G (2006). "Analyses comparatives de productions écrites d'apprenants de français et de locuteurs francophones à l'aide d'outils d'extraction automatique du langage", *revue ALSIC (Apprentissage des Langues et Systèmes d'Information et de Communication)*, vol. 8, 2006, pp. 81-94. http://alsic.u-strasbg.fr

Avouris, N., Komis, V., Margaritis, M.& Fiotakis, G. (2004). "An environment for studying collaborative learning activities". *Journal of International Forum of Educational Technology & Society , Special Issue on Technology – Enhanced Learning*. Vol. 7, 2. pp. 34-41.

Baldry, A. & Thibault, P.J. (2006). *Multimodal Transcription and text Analysis*. Equinox : Londres.

Baron, G-L., Bruillard, E. & Sidir, M. (Dir.) (2005). Symposium Symfonic "*formation et nouveaux instruments de communication*. Amiens, janvier : Université de Picardie. http://archive-edutice.ccsd.cnrs.fr/edutice-00000897

Berlin (2003). *Appel de Berlin d´octobre 2003 sur "Open Access to Knowledge in the Sciences and Humanities"*. Institut Max Planck : Munich. www.zim.mpg.de/openaccess-berlin/berlindeclaration.html

Betbeder, M.-L., Tissot, R. & Reffay, C. (2007). " Recherche de patterns dans un corpus d'actions multimodales". Conférence EIAH'2007 (Environnements Informatique pour l'Apprentissage Humain), Lausanne, CH., juin, 12 p.

Betbeder, M.-L., Reffay, C. & Chanier, T. (2006). "Environnement audiographique synchrone : recueil et transcription pour l'analyse des interactions multimodales". *JOCAIR 2006, JOurnée Communication et Apprentissage Instrumentés en Réseau*, Amiens, juillet. 15 p.

Betbeder, M-L. & Tchounikine, P. (2005). "Conception d'activités collectives dans un contexte d'apprentissage". In Teulier, R., Charlet, Tchounikine, P. (dir). *Ingénierie des Connaissances*. Paris : L'Harmattan. p.437-458.

Blache, P., Bonastre, J.F. & Nguyen, N. (2004). « Traitement de l'écrit et de la parole », *Traitement automatique des langues (TAL)*, vol. 45,3.

Chanier, T. (2004) *Archives ouvertes et publication scientifique. Comment mettre en place l'accès libre aux résultats de la recherche* ? Paris : L'Harmattan. 186p. ISBN : 2 7475 7695 7. http://archivesic.ccsd.cnrs.fr/sic_00001103

Chanier, T. & Vetter, A. (2006). "Multimodalité et expression en langue étrangère dans une plate-forme audio-synchrone". *Apprentissage des langues et Système d'Information et de Communication (Alsic)*, vol. 9. http://alsic.org. 20p. http://alsic.u-strasbg.fr/v09/chanier/alsic_v09_08-rec3.htm

Chanier, T., Vetter, A., Betbeder, M.-L. & Reffay, C. (2006). "Retrouver le chemin de la parole en environnement audio-graphique synchrone". In Mangenot, F. & Dejean-Thircuir, C. (Dir.) *Les échanges en ligne dans l'apprentissage et la formation, numéro thématique Le Français Dans Le Monde*, juillet.

Chapelle, C.A. (2004). "Technology and second language learning: expanding methods and agendas. *System,* 2004, pp. 593-601.

Crego, J. (2004). Conception collaborative d'un espace numérique de travail. Mémoire de DEA, université Lyon 2.

Daille, B & Romary, L. (dirs) (2001). *Linguistique de corpus. Traitement automatique des langues (TAL*), vol. 42,2.

Symposium Epal

Djouad, T. (2005). TASync : un outil de transcription et de visualisation des actions multimodales. Rapport de master, Laboratoire d'Informatique de l'Université de Franche-Comté, août 2005.

Ganascia, J.-G. et Bourdaillet, J. (2006). "Alignements unilingues avec MEDITE". *JADT*, *Huitièmes Journées Internationales d'Analyse Statistique des Données Textuelles*, Besançon, avril.

Granger, S. Hung, J. & Petch-Tyson,S. (2002). *Computer Learner Corpora, second language acquisition and foreing language teaching*. John Benjamins Publishing : Amsterdam.

Greffier, F., Reffay, C. (2006). "Le forum de discussion, un outil opérationnel pour réussir en FAD ? ", *JOCAIR'06 JOurnée Communication et Apprentissage Instrumentés en Réseau*, Amiens, juillet. 15 p.

Greffier, F. (2005). "Le tutorat, un geste pédagogique". *Tutorat et logiques industrielles*. *Revue Distances et Savoirs*, vol. 3, 2, pp 231-250.

Henri, F. & Charlier, B. (2005). "L'analyse des forums de discussion Pour sortir de l'impasse". In Baron G-L., Bruillard E., Sidir M. (Dir.)

Herring, S.C (2004). "Computer Mediated Discourse Analysis: An approach to researching online behaviour". In Barab, S.A., Kling, R. & Gray, J.H. *Designing for virtual communities in the service of learning*. Cambridge University Press. pp. 338-376.

IMS (2003). *IMS Learning Design Information Model* Version 1.0 Final Specification, 20 January 2003. IMS Global Learning Consortium : Burlington, MA. www.imsglobal.org/learningdesign/ldv1p0/imsld_infov1p0.html.

ISGS (2005). *Second Congress of the International Society for Gesture Studies*, Lyon (France). http://vislab.cs.vt.edu/~gesture/multimodal/workshop/index.html

Jeannot, L., Vetter, A. & Chanier, T. (à paraître 2006). "Repérage des stratégies des apprenants et du tuteur dans un environnement audio-graphique synchrone". In Mangenot, F. & Dejean-Thircuir, C. (Dir.) *Les échanges en ligne dans l'apprentissage et la formation, numéro thématique Le Français Dans Le Monde,* juillet.

Jouis, C, (1995). "SEEK, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe"**,** *Actes des 6ème Journées Acquisition, Validation, (JAVA 95)*, INRIA et AFIA, pp. 159-172, Grenoble, avril.

Kern, R., Ware, P. & Warshauer, M. (2004). "Crossing frontiers: new directions in online pedagogy and research". *Annual Review of Applied Linguistics*, Vol. 24. pp. 243-260.

Kress, G. & Jewitt, C. (2001). *Multimodal teaching and learning: the rhetorics of the science classroom*. Continuum : Londres.

Lamy, M-N. (2006). "Interactive Task Design and the Whole Learner" In Garcia Mayo, P. (ed.) *Investigating Tasks in Formal Language Settings*, Multilingual Matters.

Lamy, M-N. (2006). "Conversations multimodales : l'enseignement apprentissage de l'oral à l'heure des écrans partagés" In Mangenot, F. & Dejean-Thircuir, C. (eds.) *Les échanges en ligne dans l'apprentissage et la formation, Numéro spécial, Le Français Dans Le Monde,* juillet.

Lamy, M-N (2005). "Apprenants et conférences électroniques : facilitation et détournements". *Proceedings of the conference 'L'institution face au numérique', SIF (Séminaire Informatique et Formation)*, Maison des Sciences de l'Homme, Paris, December 2005. http://sif2005.mshparisnord.org/pdf/lamy.pdf

Lamy, M-N. (2004). "Oral Conversations Online: Redefining Oral Competence in Synchronous Environments", in *ReCALL*, Vol 16, n° 2.

Lamy, M-N. & Hassan, X.P. (2003). "What influences reflective interaction in distance peer learning? Evidence from four long-term online learners of French". *Open Learning Journal*, Vol 18, n° 1, pp 39-59.

Lejeune, A. & Marino, O. (2005). "IMS Learning Design, un langage de modélisation pédagogique", Cours de la Troisième Ecole thématique du CNRS sur les EIAH "Modèles, architectures logicielles et normes pour le développement et l'intégration des EIAH" http://ecole-cnrs.univ-lemans.fr/eiah2005/fichiers/cours/Cours4.pdf

Levine, P. & Scollon, R. (2004). *Discourse & technology: multimodal discourse analysis*. Georgetown University Press : Washington, DC.

Mbala, A., Reffay, C. & Chanier, T. (2003). "SIGFAD : un système multi-agents pour soutenir les utilisateurs en formation à distance". *Actes de la conférence Environnements Informatiques pour l'Apprentissage Humain (EIAH'2003)*, Strasbourg, France, avril. Pp. 319-330.

Mbala, A., Reffay, C. & Chanier, T. (2002). "Integration of automatic tools for displaying interaction data in computer environments for distance learning". In S.A. Cerri, G. Guardères, and F. Paraguaçu (dir.) *Proceedings of Intelligent Tutoring System conference (ITS'02)*. Springer-Verlag. pp 841-850.

Mondada, L. (2003). Téléchirurgie et nouvelles pratiques professionnelles : les enjeux interactionnels d'opérations chirurgicales réalisées par visioconférence, Sciences Sociales et Santé.

Noras, M., Reffay, C. & Betbeder, M.-L. (2007). " Structuration de corpus de formation en ligne en vue de leur échange", Conférence EIAH'2007 (Environnements Informatique pour l'Apprentissage Humain), 6 p., Lausanne, CH., juin 2007.

Symposium Epal

Noras, M. (2006). "Un besoin de spécifications des corpus de formation en ligne", *1ères rencontres Jeunes Chercheurs en Environnement Informatique sur l'Apprentissage Humain 2006 (RJC-EIAH'06)*, Evry, 11-12 mai.

OAI-MPH (2002). *The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14*. Document Version 2002/07/05. Open Archive Initiative. www.openarchives.org/OAI/openarchivesprotocol.html

Reffay, C., Noras, M., Chanier, T. & Betbeder, M.-L., (2007), " Contribution à la structuration de corpus de formations en ligne pour un meilleur partage en recherche", Communication au colloque EPAL (Echanger Pour Apprendre en Ligne), Grenoble, France, juin 2007.

Reffay, C. & Teutsch, P., (2007), "Anonymisation de corpus réutilisables", Poster et démonstration, Conférence EIAH'2007 (Environnements Informatique pour l'Apprentissage Humain), Lausanne, CH., juin 2007.

Reffay, C. & Lancieri, L., (2006), "Quand l'analyse quantitative fait parler les forums de discussion", Revue STICEF, Volume 13, 2006, ISSN : 1764-7223, mis en ligne le 02/03/2007, http://sticef.org

Reffay, C. & Betbeder, M.-L. (2006). "Spécificités des plates-formes audio-synchrones pour un dispositif de formation". In Mangenot, F. & Dejean-Thircuir, C. (Dir.) *Les échanges en ligne dans l'apprentissage et la formation*, numéro thématique *Le Français Dans Le Monde*, juillet.

Reffay, C. (2005). "Réseaux sociaux et analyse de traces des forums d'une communauté d'apprentissage". In G.-L. Baron, E. Bruillard, and M. Sidir (Dir.).

Reffay, C. & Chanier, T. (2003). "How social network analysis can help to measure cohesion in collaborative distance-learning", *Proceeding of Computer Supported Collaborative Learning conference (CSCL'2003), June, Bergen, Norway.* Kluwer Acedemic Publisher. pp 343-352.

Reffay, C. & Chanier, T. (2001). "CUMULI : construction d'une mémoire du groupe dans l'interaction en FAD". *Sciences et Techniques éducatives (STE)*, vol. 8. pp. 155-158.

Rohlfing, K., Loehr, D., Duncan et al. (2005). "Comparison of multimodal annotation tools – workshop report". *Second Congress of the International Society for Gesture Studies*, Lyon. 15-18 Juin 2005. http://vislab.cs.vt.edu/~gesture/multimodal/workshop/Report.pdf

Salmon-Alt, Romary, L. & Pierrel, J.-M. (2004). "Un modèle générique d'organisation des corpus en ligne". *Traitement automatique du langage (Tal)*, vol. 45, 3. pp. 145-169.

Smtih, B. & Gorsuch, G.J. (2004). Synchronous computer mediated communication captured by usability lab technologies: new interpretations. In *System* (2004). pp. 553-575.

SSHRC (2005). *Research Data Archiving Policy* Conseil de recherche en sciences humaines du Canada. http://www.sshrc.ca/web/about/policies/edata_e.asp

System (2004). *Incorporating multimedia capability in the reporting of applied linguistics research. Numéro thématique, revue System*, Vol. 32.

Tribble, C. & Jones, G (1997). *Concordances In The Classroom : a resource book for teachers*. Houston : Athelstan.

Valcke, M. & Martens, R. (2006). *Methodological Issues in Researching CSCL , Special issue of Computers & Education*, Vol. 46, 1. pp. 1-104.

Véronis, J. (2004). *Le traitement automatique des corpus oraux. Traitement automatique des langues (TAL)*, vol. 45,2.

Sites Internet

ATONET (2006). Réseau pour l'échange de ressources et de méthodologies en analyse de texte assistée par ordinateur. http://www.atonet.net

CES (2000). Site du *Corpus Encoding Standard*. http://www.cs.vassar.edu/CES/

COCOSDA (2002). *Outils et formats pour l'encodage de corpus*. Projet Cocosda (Co-ordination and Standardisation of Speech Databases and Assesment Techniques). http://www.ldc.upenn.edu/annotation/

CLAPI (2006) : Corpus de Langues Parlées en Interaction. ICAR - groupe ICOR, Université le Lyon 2, http://clapi.univ-lyon2.fr/

Master FOAD (2006). Site du *master FOAD (formation ouverte et à distance)*. Université de Franche-Comté. http://masterfoad.univ-fcomte.fr

OLAC (2006). Site de *Open Language Archives Community.* http://www.language-archives.org/

TALKBANK (2006) : http://talkbank.org/

TC37/SC4 (2006). Site du *groupe de travail ISO "Language Resources Management".* http://www.tc37sc4.org

Symposium Epal

TEI (2006). Site du *Text Encoding Initiative*. http://www.tei-c.org

Logiciels

ANVIL version 4.5 (2003). Consulté en juin 2007: http://www.dfki.uni-sb.de/~kipp/anvil/

CLAN (Computerized Language Analysis). Consulté en juin 2007: http://childes.psy.cmu.edu/

ColAT Collaboration Analyse Toolkit (HCI Group : http://hci.ece.upatras.gr/index.php?lang=iso-8859-1) http://hci.ece.upatras.gr/index.php?option=com_content&task=view&id=101&Itemid=103

ELAN version 2.4.1 (EUDICO Linguistic Anotator). Consulté en juin 2007: http://www.mpi.nl/tools/elan.html

EXMARaLDA version 1.3.2 (Extensible Markup Language for Discourse Annotation). Consulté en juin 2007 : http://www1.uni-hamburg.de/exmaralda/index-en.html

MacVisSTA version 2 (Macintosh Visualization for Situated Temporal Analysis). Consulté en juin 2007 : http://vislab.cs.vt.edu/~rtr/

MOT/MOT+ : Consulté en juin 2007. http://www.licef.teluq.uquebec.ca/fr/realisations/mot1.htm

Praat. Consulté en juin 2007: http://www.praat.org

Reload : Consulté en juin 2007. http://www.reload.ac.uk/

TasX : Consulté en juin 2007: http://medien.informatik.fh-fulda.de/tasxforce/TASX-annotator/index.html

Symposium Epal

Référence de cet article : N. Avouris, G. Kahrimanis, G. Fiotakis, E. Voyiatzaki, M. Margaritis. "Tools supporting Collaborative Learning and Interaction Analysis: Synergo and ActivityLens". Composition from previous publications of the HCI Group, Patras university, adapté pour le symposium EPAL\Corpus, Grenoble, 7 juin 2007.

Tools supporting Collaborative Learning and Interaction Analysis: Synergo and ActivityLens.

N.Avouris, G. Kahrimanis, G. Fiotakis, E. Voyiatzaki, M. Margaritis

University of Patras, Greece (http://hci.ece.upatras.gr)

1. Introduction

This document[1] presents two tools designed and developed by the Human-Computer Interaction Group of the University of Patras to support collaborative learning and collaborative interaction analysis. These are *Synergo* and *ActivityLens*. Synergo constitutes a synchronous collaboration support environment that provides integrated tools for analysis of interaction. A large corpus concerning online human interactions is generated and further analysed using Synergo. ActivityLens constitutes an environment that aids analysis of collaborative learning activities using multiple sources of data such as sequential logs, video captures and documents related to the activity, a corpus of oral interactions can be analysed using this tool, and for this reason is considered relevant to generation and analysis of online human interactions.

In following, a presentation of the Synergo environment is included accompanied by a study on the use of the tool. Next, ActivityLens as an analysis tool that extends the analysis facilities of Synergo is included, focusing on the manipulation of various sources of analysis data and on the support for qualitative methodological studies. In the third part the context of current use of the two tools is described, mentioning cases of research groups that have used the tools. Finally, this document ends with a discussion on standards on representation of interaction data and the related emergent need to develop widely adopted descriptions of corpora of interaction analysis data and the potential benefits of that approach to the research community in the future, based on some concrete examples from the experience of the Unvierstity of Patras HCI Group and other research groups.

Symposium EPAL

2. Collaborative Problem Solving through Synergo

Synergo is a collaboration support environment based on the Abstract Collaborative Applications Building Framework (ACABF). This underlying framework has also been used for building ModellingSpace (Margaritis et al. 2004) and ModelsCreator v3 (Fidas et al. 2002). Synergo architecture supports synchronous collaboration, as well as integration of collaboration analysis and visualization tools.

The Synergo environment that can be downloaded from the web site of the University of Patras HCI Group[2] is a client-server distributed application, which comprises a suite of interconnected tools to support collaborative activities supported by textual collaboration tool and a shared workspace in which various diagrammatic representations can be collaboratively built. The architecture of a set of workstations in which Synergo is used is shown in figure 1. In this figure a typical Synergo classroom is shown, which could be made of students who are collocated or are

[1] This document contains material from previous publications of the HCI group and is used as handouts for a workshop on EPAL\Symposium with the subject: Designing, re-using and exchanging online learner corpora, Grenoble, June 7th, 2007. (http://mulce.univ-fcomte.fr/epal_symposium/index_english.html) .

[2] http://hci.ece.upatras.gr/synergo

dispersed and communicate over the Internet. One workstation is set up as Synergo server while the others are Synergo clients which are set up to recognize their specific Server.

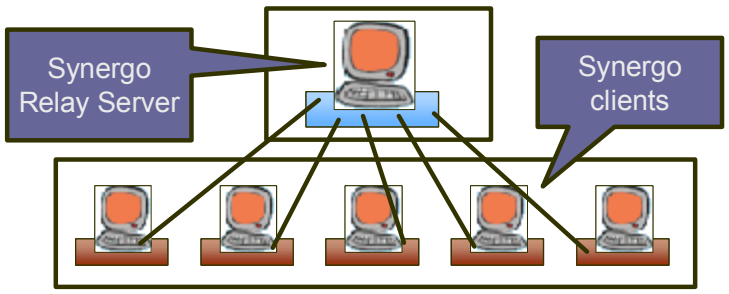

Figure 1. Synergo typical setting

The Synergo software can be downloaded from the Synergo site and can be installed on a workstation. This is a Java application of around 2 MB that can run if the Java run time environment JDK 1.5 or JDK 1.6 has been installed on the specific workstation. A separate file which includes the java virtual machine can be downloaded, which is of much larger size (around 16 MB), if in doubt about the Java run time environment already installed. Finally the possibility of running Synergo using Java web start technology is made available, however this latter case, which serves for automatic updates of the software may become unstable if later versions of Java runtime environment are installed on the specific workstation. Currently (May 2007) version 3.01 of Synergo is available for free download.
A Synergo portal has been created on which one can find user and administrator manuals and replies to frequently asked questions and can subscibe to receive news. The intention is in the future to include in this Synergo portal material from users of Synergo in the form of logfiles that can be compatible to a common standard for metadata and annotations. In figure 2 one can see the home page of the Synergo portal.

Symposium Epal

Figure 2. Home page of the Synergo Community portal

Once the Synergo software is installed on a workstation, one may decide if the particular workstation would be used as server or as a client. For the client installations, the server should be defined through a dialogue in the set up menu, shown in figure 3.

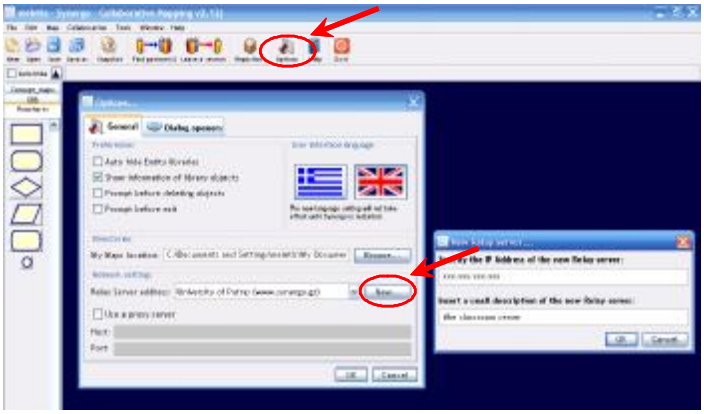

Figure 3. Set up of the Relay server of the client

In a typical collaborative problem solving situation, the coordinator will arrange the groups of collaborating partners that would work together on a given problem. There is a facility through the Relay Server to set up groups (see figure 4) or an individual partner can start a session selecting one or more other users who are on line.

Figure 4 Creation of groups of students. In this case a random group creation algorithm is used, defining only the group size.

When the students start a session find themselves in front of an environment like the one shown in figure 5. The central window is the shared workspace in which they can build collaboratively a shared diagrammatic representation out of primitive objects. Synergo supports building of different kinds of diagrams. It contains libraries for building flowcharts, entity-relationship diagrams, concept maps, data flow diagrams etc. On the left-hand side column of figure 5, libraries of primitive objects are shown. On the right had side a chat tool is used for online communication in textual form, while an awareness tool provides feedback on the state of the partners, for instance if the partner is not focussing on the Synergo window is indicated as the partner being idle. So in general the activity in such an environment involves both gestural activity in the workspace and verbal activity through the chat tool. The log file of the activity contains an ordered set of actions that may be interpreted in the context in which they take place, as they are characterized by elliptic nature are related to the state of the workspace.

Synchronous collaboration using a workspace is a case of computer-supported collaboration based on the concept of shared artefact (Dix et al, 1998). The related

notion of feed-through this artefact (the diagrammatic representation) implies that one participant's manipulation of shared objects can be observed by the other participants. This communication through the artefact can be as important as direct communication between participants. Considering that the collaborative activity is done mainly between partners at a distance, the direct communication mechanism plays also an important role, complementary to the gestural activity. A text communication has been used in Synergo, as it was found beneficial for learning activities, in which textual expression involves higher cognitive skills especially for argumentation tasks, often required in collaborative problem solving activities.

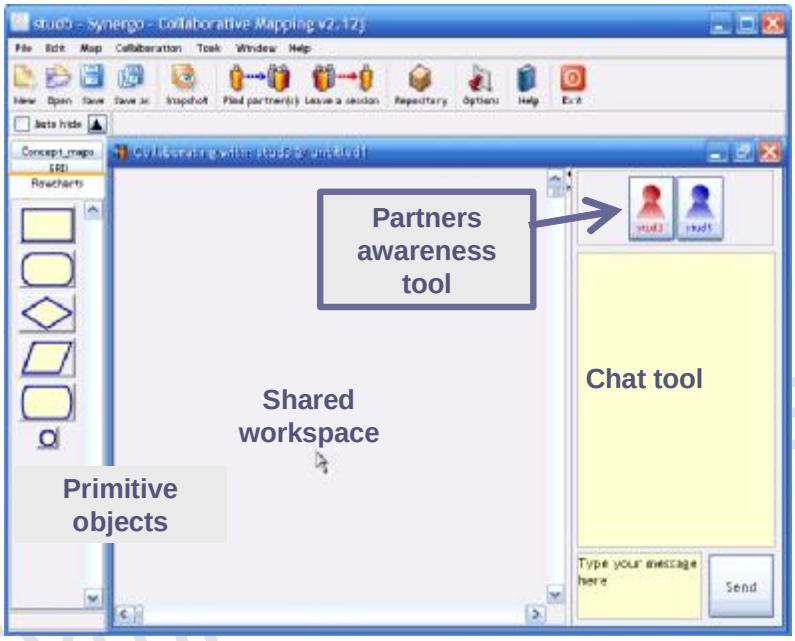

Figure 5. Typical Synergo client environment

The degree of synchronization between the client workstations in Synergo varies. A strict WYSIWIS (what you see is what I see) is implemented in the shared problem-solving window. This is because most of communication and reasoning is based on this shared viewpoint, which becomes the main grounding mechanism of dialogue and through which eventually common understanding can occur. However all additional operations outside this shared workspace, e.g. relating to private not-shared windows, saving of the diagram may be performed independently by partners involved, a model-level coupling approach according to Suthers (2001). This approach, also known as relaxed WISIWYS, guarantees only that users will see the same state of a shared model, but the views and their overall experiences may be entirely different, something which is also attributed to the different characteristics of the workstations with which they interact, in terms of screen resolution, screen size and other environmental conditions. It should be added that the typical Synergo activities usually involve interleaving of private and collaborative tasks, e.g. first look at the problem and experiment individually and then discuss your view as a group. This is usually implemented through multiple windows, some of which are shared and some private, see figure 6. The implication for the logfile that is produced out of an activity is that it contains a partial view over the user experience as it is related just to the shared workspace and dialogue and not to the overall user activity monitoring.

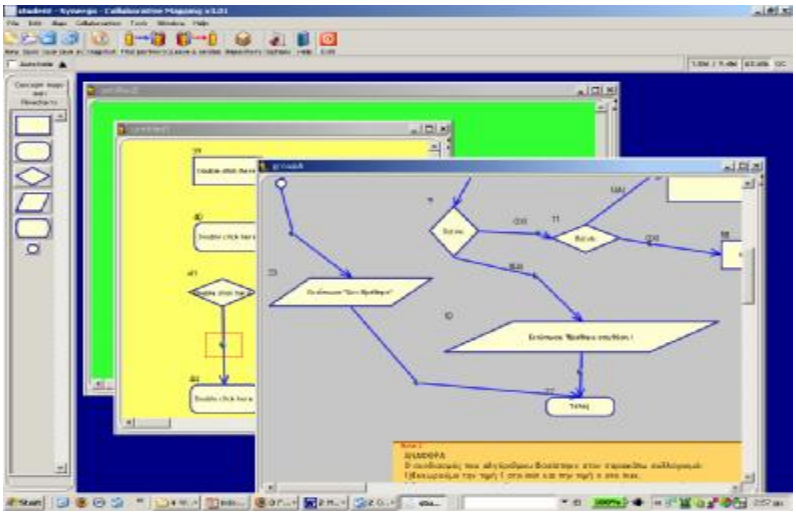

Figure 6. Synergo multi-window interface

An additional important characteristic is the support for coordination of user activity, in particular with relation to the shared activity space. In Synergo, a floor control coordination mechanism is included. This mechanism involves the notion of the *Action Enabling Key*, which is owned by one of the participants at any given time. This key owner can then act in the shared workspace, while the rest just observe this activity and make comments through the chat tool. This mechanism is supported by key request, key accept, key pass, key reject functions, which can be found in the Coordination Panel. Experiments with this floor control mechanism, see also (Fidas et al. 2002) and (Komis et al. 2002), demonstrate that it supports reasoning about action, as partners need to reason and negotiate during key requests. Synergo users may opt for this mechanism or may decide to act in the shared activity space with no specific floor control, in which case locking is effected at the level of the single entity.

In the frame of the collaborative use of Synergo, a dialogue tool has been integrated, shown at the right hand side panel of figure 5 (chat tool), which is based on an instant messaging protocol, using the same point-to-point connection and protocol of the shared activity space. Through this, text messages are exchanged during collaborative problem solving. The chat tool, is activated from the collaboration panel. The possibility of definition of dialogue openers is also included in this tool, as shown in figure 7.

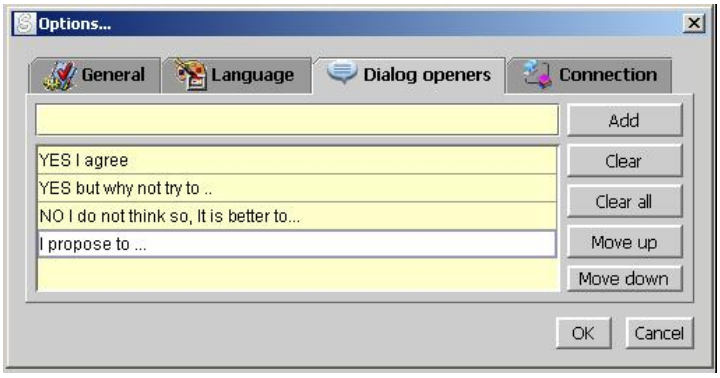

Figure 7. Examples of dialogue openers of the chat tool

Symposium Epal

Other means for exchange of text messages are the *sticky notes* as text containers positioned in the activity space, associated to existing objects, through which, pointing gestures (deixis) can be simulated.

2.1 Analysis of Synergo interaction data

The activity in Synergo is monitored and logfiles are generated and made available for inspection by the users or researchers. An example of an annotated logfile is shown in figure 8. In this extract two partners discuss the development of a flow chart. Some snapshots of the state of the activity space are included by the analyst in order to provide some background information on the context of the activity.

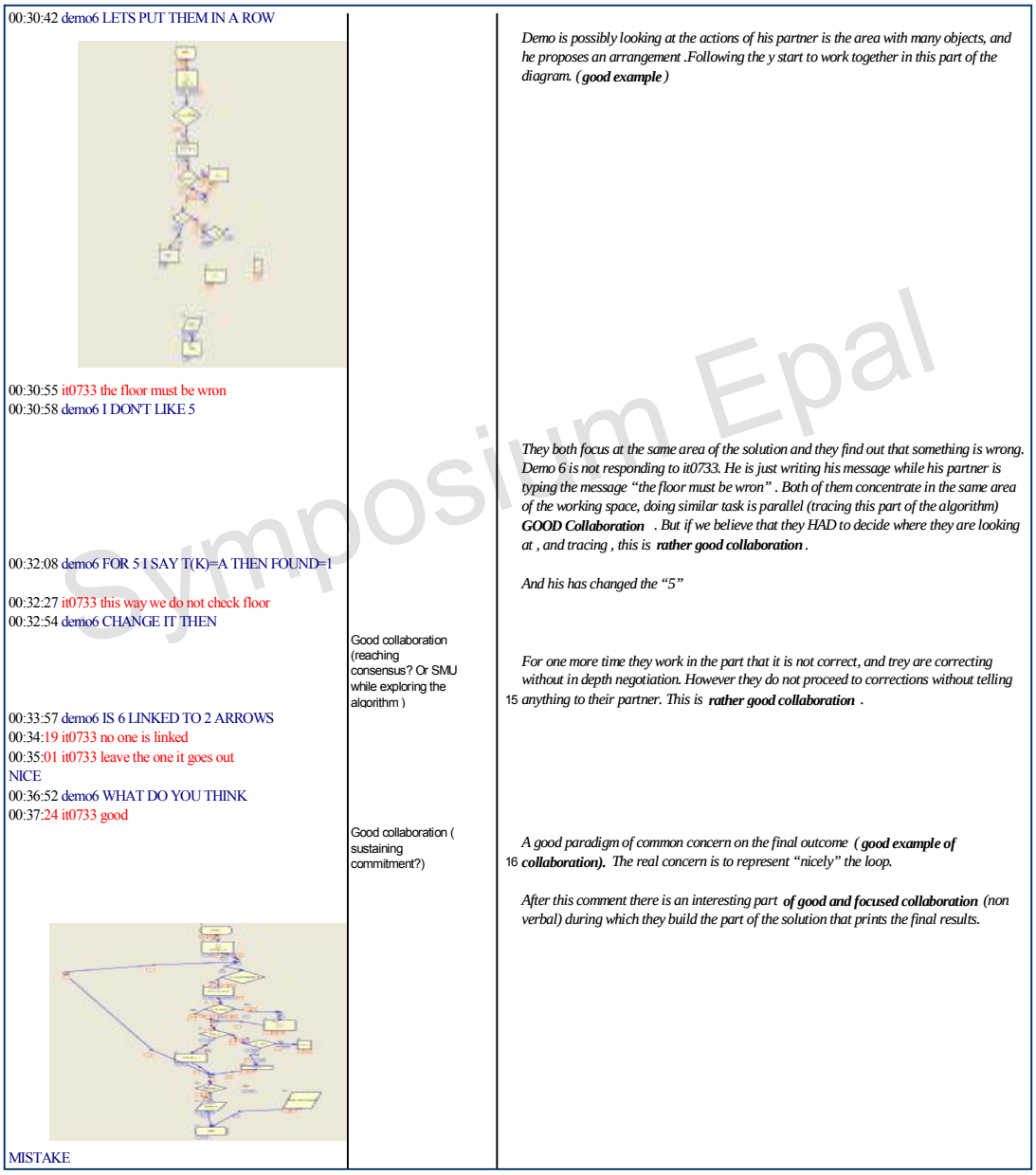

Figure 8 An extract of an annotated logfile

Symposium Epal

An important feature of Synergo relates to integrated tools for analysis of the generated logfiles. A set of Analysis and Visualization tools are included in the Synergo environment. These are mainly used by the teachers and researchers, while limited versions of the tools may be used in some cases by the students themselves as meta-cognitive aids, as is the case of the level of collaboration monitoring display. The main functionality of the Synergo Analysis tool is the presentation and processing of logfiles, which are created during Synergo use. These logfiles contain actions and text messages of all partners, in sequential order. The logfiles are based on the format of the coordination and communication protocol and are stored in XML. These files can be viewed, commended and annotated by the researchers, using an adequate analysis framework, as discussed by Avouris et al. (2003). A related functionality is the capability of the analysis environment of posterior reproduction of the modelling activity, using this logfile, in a step-by-step or continuous way (playback tool). This is complementary to the logfile inspection and annotation functionality.

In this section we describe the key parameters through which we can model collaborative problem solving activity in Synergo. This model influences the way that the logfiles are structured and historical data of use of this collaborative environment are annotated.

We suppose that the activity involves a small group of subjects (actors) who are engaged in collaborative problem solving (2 to 5 actors). Problem solving activity is usually considered as a process of refinement of abstract ideas ("abstract objects") and externalisation of these ideas in the form of parts of the solution to the given problem. Collaborative activity is based on communication, which takes the form of either direct communication acts or operations in the shared activity space. The activity is defined according to the following four dimensions:

- The time dimension t : (when the action is taking place)
- The actors' dimension: $A=\{A_1, A_2,..., A_k\}$ (who is acting)
- The objects' dimension: $O=\{O_1, O_2,..., O_1\}$ (the object of action in shared space):
- The typology of events dimension: Ty (what is the type of action).

This latter dimension leads to interpretation of the activity that takes place. It is assumed that there is an existing analytical framework, which defines this typology Ty. If r is the finite number of expected event types, then we define a set $Ty=\{T_1, T_2,..., T_r\}$ as the analytical framework of the study. Ty can be defined by the framework user.

Using the above four dimensions we can describe any given activity as a set of discrete non-trivial events produced by the actors, contained in the logfile. These define an ordered set of m events $E=\{E_1, E_2,..., E_m\}$. Each one of these events is related to meaningful actions of the actors who interact with objects of set O incrementally contributing to the problem solving activity. Each event is defined as a tuple $E_{i_{tAOT}}=(t_t, A_A, [O_O], [T_T])_i$ where $i\in[1,m]$, t the event timestamp, A the actor who performed the action of the specific event, O an optional parameter referring to the object of the specific action and T an optional parameter which interprets the event according to the analysis framework Ty.

This is a useful model for ethnographic kind of studies, as it describes the activity in a phenomenological level (occurred communication or gestural actions). Every time an event is produced by the actors, this is recorded and a history of such events, i.e. an ordered list of Es can be produced, as a result of such an activity. No mental,

Symposium EPAL

cognitive or intentionality operators are defined, as these can be generated later as interpretations of the recorded activity. This model permits further analysis and interpretation of the activity, while quantitative indices of the activity can be easily produced or visualizations can be generated (Margaritis et al. 2004).
Synergo adheres to a typology of generated events, thus automating the task of categorization of observed events (insertion, modification, deletion of primitive objects in the shared space and exchange of text messages), every time such an operation is recorded, this is automatically categorized according to the scheme of analysis defined by the user. OCAF suggests interpretation of exchanged messages (written dialogues during collaboration by distance), or recorded oral utterances (during face to face collaboration), in relation to operations towards "objects" of the activity space, using a language for action approach (Winograd 1987), defining a unifying framework for analysis of dialogue and action.

2.2 A case study of analysis of collaboration with Synergo

In this section we describe an example of a study that involved analysis of collaborative activity using the Synergo tool. The activity involved building of a concept map of an Internet service (an electronic bookshop was chosen as the example of the service to be model by the participants in this case) by small groups of students of an undergraduate University course, in the frame of one lab session (45'). We focus on one of these groups made of 4 students in this section. The logfile of the activity of this specific group was studied using the Synergo. More details of this study can be found in Avouris et al. (2004b). First the relative weights of the activity types and the actors were defined, as seen in figure 9(a) In our case events related to creation and modification of sticky notes are assigned lower weight (0.3), as they are used for administration issues.

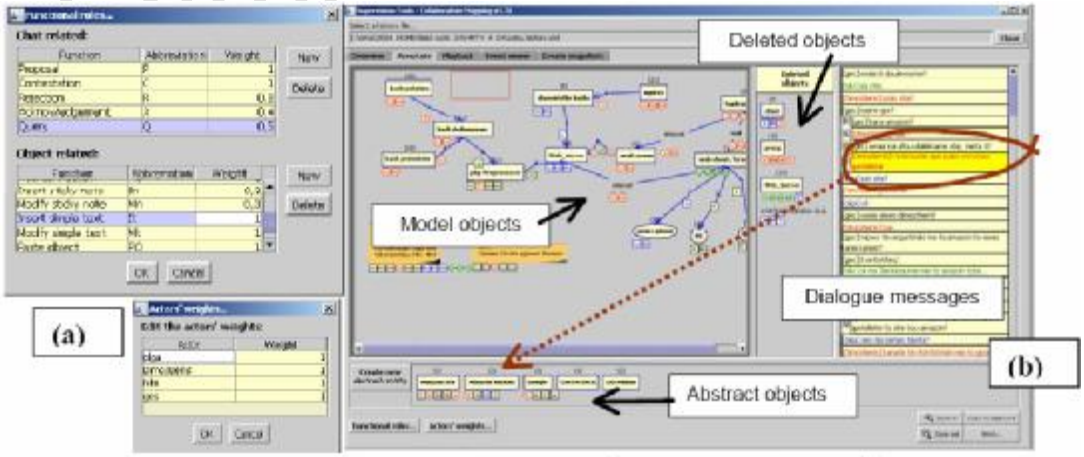

Figure 9. (a) Definition of activity type W(t) and Actors weights W(A) and (b) annotation of dialogue events

Subsequently the dialogue events were annotated according to the defined typology. This phase involved definition of abstract entities that appeared in the dialogue. The dialogue annotation window is shown in figure 9(b). Three types of objects are shown in this window: the components of the final solution in the main panel (model objects), the deleted components in the vertical panel and the abstract

components at the bottom panel. In the example of figure 9(b) a dialogue event is associated to the abstract object “Amazon model”: Actor Ges said: ”what to assign to the Amazon site?”, This dialogue message was categorized as a Q (Query) and was associated to the abstract object “Amazon model”, by a simple drag operation.
After annotating dialogue events, we are able to playback the activity and produce in numeric and visual form the evolution of the Collaboration Factor, a metric developed that provides a quantitative approach of the rate of collaboration (Avouris, et. al., 2004c). This is shown in figure 10(a). Some other indices, like the density of actors’ activity of various types in the shared activity space, can be produced automatically, from the Synergo logfile. Also the contribution of each actor in the activity can be visualized. In figure 10(b) the actor contribution of “insert object” events and chat messages is shown. Each line of these diagrams represents one of the four group members. From this picture, it is deduced that the second actor shows relatively low activity. More complex indices, like the Collaboration Factor mentioned above, are produced as a result of interpretation of actions and dialogue events. An example is the visual representation shown in figure 10(a).

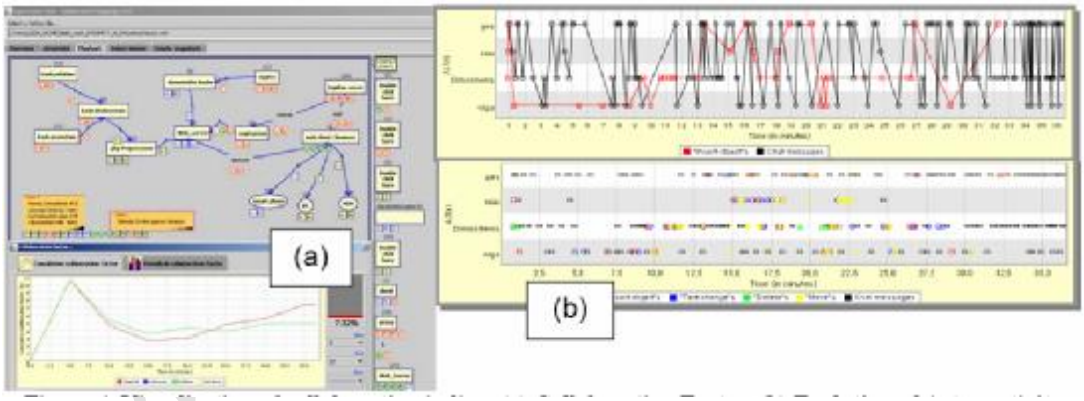

Figure 10. Visualization of collaboration indices (a) Collaboration Factor, (b) Evolution of Actor activity

This provides an indication of the degree of collaboration of the group of the four students as they are building the e-shop concept map. From this graph it seems that while (for the first period of the activity) the degree of collaboration was high, subsequently the partners became more individualistic, working on parts of the solution, as also shown in the annotated concept map of figure 4(a). Later on towards the end of the session, there is more interaction, at the level of specific concepts and entities.

3. Interrelation of activity logfile to other behavioural data through the ActivityLens

It should be observed that structured data, like a typical logfile discussed in previous sections, takes usually the form of an ordered list of events that were recorded at the user interface of a software environment, like Synergo. It contains a record of the activity of one or more actors, from the rather restrictive point of view of their fingertip actions. However a lot of contextual information relating to the activity, as well as results of the

activity in print or other forms, oral communication among the actors, is not captured through this medium. Thus, in this section we present an analysis environment that was developed by the HCI Group in order to permit integration of multiple media collected during collaborative activities and to allow the application of qualitative methodologies that relate to the fields of ethnomethodology (Garfinkel, 1967), conversation analysis (Edwards, & Potter, 1992), interaction analysis (Jordan, & Henderson, 1995), video analysis (Heath, 1986) and ethnography (Hammersley, 1982) are applied.
ActivityLens is the name given to the new version of the environment previously known as *Collaboration Analysis Tool (ColAT).*
ActivityLens[3] is the environment that is used for building an interpretative model of the activity in the form of a multilevel structure, following an Activity Theory approach (Bertelsen & Bodker, 2003), incorporating pointers and viewers of various media. ActivityLens permits fusion of multiple data by interrelating them through the concept of the universal activity time. Figure 11 shows the dialogue of creation of a new analysis project and inter-relation of multiple sources of data. The analysis process during this phase, involves interpretation and annotation of the collected data, which takes the form of a multilevel description of the activity.

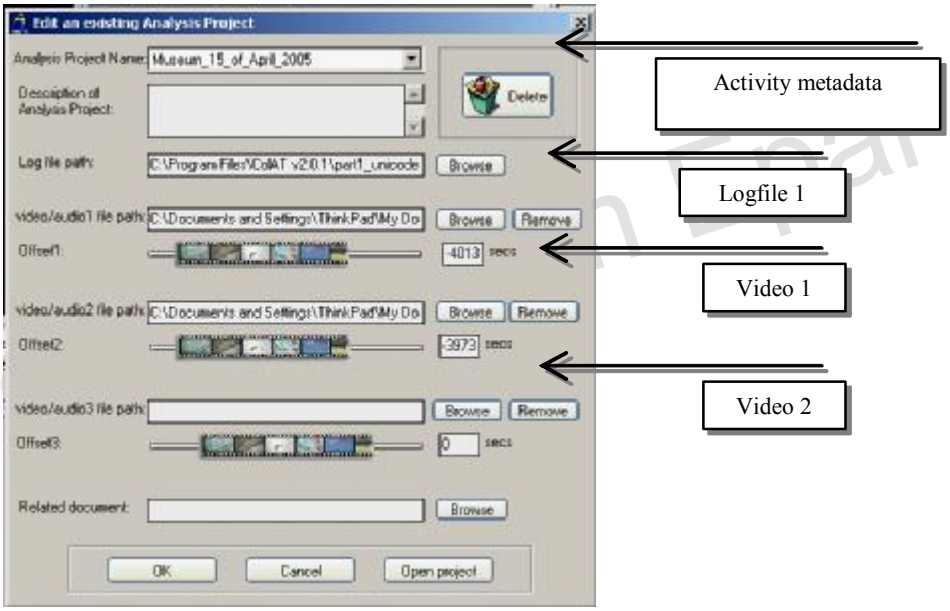

Figure 11. The ActivityLens environment: Project definition in which multiple logfiles and video/audio sources are synchronized by defining their corresponding time offsets.

The ActivityLens tool, discussed in more detail as ColAT in (Avouris et. al., 2004b), uses the form of a theatre's scene, in which one can observe the activity by following the plot from various standpoints. The *Event view* permits study of the details of action and interaction, as recorded by a logfile, while other media like most typically video and audio recordings, capture dialogues, other behavioural data of actors (posture, gestures, facial expressions etc.), while media like screen snapshots, PDF files etc record intermediate or final outcomes of the activity. The automatically generated log of behavioural data can be expanded in two ways:

[3] ActivityLens can be downloaded from hci.ece.upatras.gr as an exe file that can be used under various versions of the Microsoft Windows operating system. The downloaded file is around 10MB

- First by introducing additional events as they are identified in the video and other media, and by associating comments and static files (results, screen snapshots etc.) to specific time stamped events.
- Second, more abstract interpretative views of the activity may be produced: the *Actions-view* permits study of purposeful sequences of actions, while the *Activity view* interprets the activity at the strategic and motivational level, where most probably decisions on collaboration and interleaving of various activities are more clearly depicted.

This three-level model is built gradually: the first level, the *Events level*, is directly associated to log files of the main events, produced and annotated, and is related through the time stamps to the media like video. The second level describes *Actions* at the actor or group level, while the third level is concerned with *motives* of either individual actors or the group.
In figure 12 the typical environment of the ActivityLens tool for creation and navigation of a multi-level annotation and the associated media is shown. The three-level model, discussed in more detail in the following, is shown on the right side of the screen, while the video/audio window is shown on the left-hand side. One other feature shown in Figure 12 is the viewer filter, through which a subset of the activity can be presented, related to specific *actors*, *tools* or *types of events*. So the logfile events related to a specific actor may be shown, or actions related to a specific tool, or a specific kind of operations.

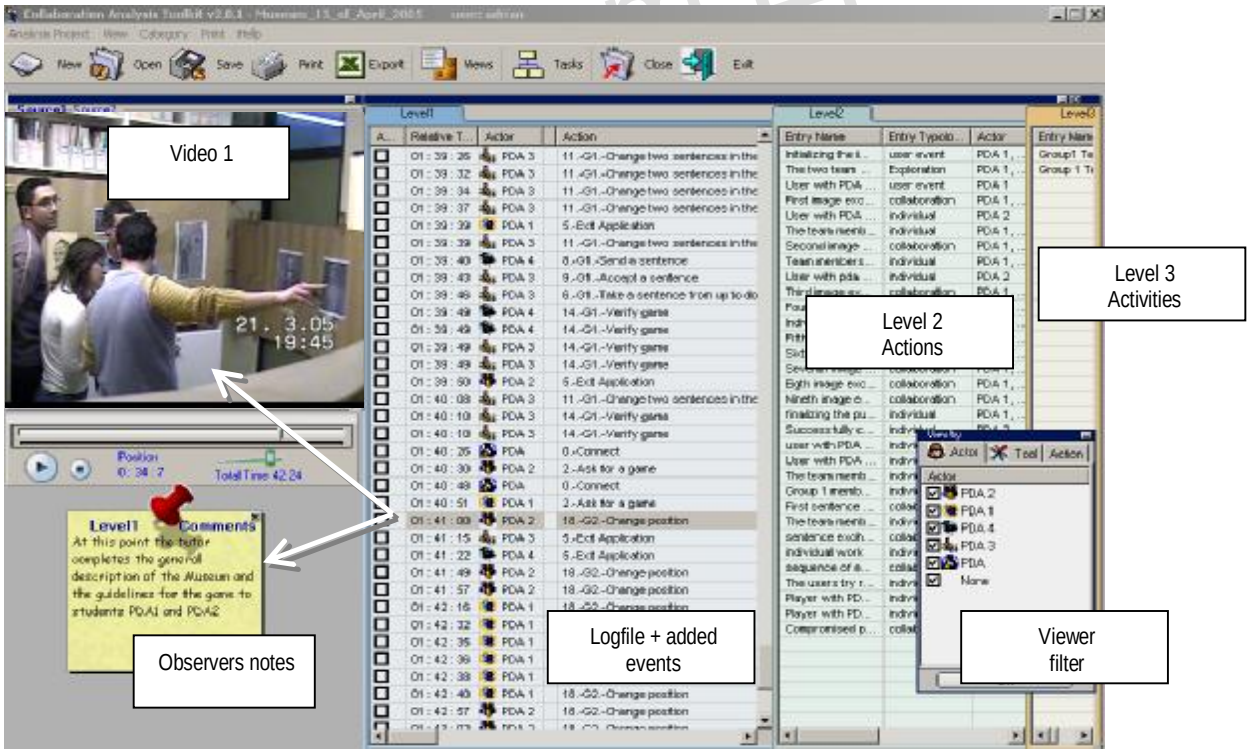

Figure 12. The ActivityLens environment: Multi-level view of problem solving activity, (The extract is from the study of Learning Activities in a Museum, discussed in Cabrera et al, 2005).

A more detailed description of the multilevel representation of the activity shown in Figure 12 is provided next. The original sequence of events contained in the logfile is shown as level 1 (*Events level*) of this multilevel model. The format of events of this level, in XML, is that produced by Synergo, ModellingSpace, CollaborativeMuseumActivity and other tools that adhere to this data interchange format (Kahrimanis et al. 2006), discussed also in section 4. Thus the output of these environments can feed into ActivityLens, as first level structure. A number of such events can be associated to an entry at the *Actions level 2*. Such an entry can have the following structure: { <ID>, <time-span>, <entry_type>, <actor)>, <comment > } where ID is a unique identity of the Action, time-span is the period of time during which the action took place, type is a classification of the entry according to a typology, defined by the researcher, followed by the actors that participated in the activity, a textual comment or attributes that are relevant to this type of action entry. Examples of entries of this level are: "Actor X inserts a link ", or "Actor Y contests the statement of Actor Z".
In a similar manner, the entries of the third level (*Activity level*) are also created. These are associated to entries of the previous *Actions level 2*. The entries of this level describe the activity at the strategy level as a sequence of interrelated goals of the actors involved or jointly decided. This is an appropriate level for description of plans, from which coordinated and collaborative activity patterns may emerge. In each of these three levels, a different event typology for annotation of the entries may be defined. This may relate to the domain of observed activity or the analysis framework used. For entries of level 1 the OCAF event typology (Avouris et al, 2003) has been used, while for the action and activity level different annotations have been proposed. In figure 13 the tools for definition of annotation scheme for actions and identity of actors and tools in ActivityLens are shown.
The various media, like video or audio that can be associated to logged events through ActivityLens can be played from any level of this multi-level model of the activity. As a result, the analyst can decide to view the activity from any level of abstraction he/she wishes, i.e. to play back the activity by driving a video stream from the *operations*, *actions* or the *activity* level. This way the developed model of the activity is directly related to the observed field events, or their interpretation.

Symposium EPAL

Figure 13. Definition of (a) tools used, (b) actors, and (c) typology of events relating each type of event to a specific color code, in ActivityLens

Other media, like still snapshots of the activity or of a solution built for a given problem, may also be associated to this multilevel model. Any such image may be associated

through a timestamp to a point in time, or a time interval, for which this image is valid. Any time the analyst requests the reproduction (playback) of relevant sequence of events, the still images appear in the relative window. This facility may be used to show the environment of various distributed users during collaboration, as well as tools and other artefacts used. Also observer comments related to events can be inserted and shown in the relevant window, as shown in the bottom left corner of Figure 12.
The possibility of viewing a process using various media (video, audio, text, logfiles, still images), from various levels of abstraction (operation, action, activity), is an innovative approach. It combines in a single environment the hierarchical analysis of a collaborative activity, as proposed by Activity Theory, to the sequential character of behavioural data.

3.1 Analysis studies with ActivityLens

The discussed tools have been used in a number of studies that involved effective analysis of collected evidence of technology-supported learning activities in various forms. Three such studies are briefly presented here.
In the study reported in (Fidas, Komis, Tzanavaris & Avouris, 2005), data were collected of groups of students interacting through the *ModelsCreator3* environment (Fidas et al. 2002). Interaction between distant group members was mediated by a chat tool while interaction between group members that were located in front of the same workstation was mainly direct conversation. Interaction in the first case was captured through the ModelsCreator3 logfile that conforms to the ActivityLens format, while the latter was captured through audio recording. By associating the two data sources, valuable information on comparison of the content of interaction that was done through the network and the dialogues of the group members was performed. The educational process was thus discussed according to various dimensions, like group synthesis, task control, content of communication, roles of the students and the effect of the tools used.
In these studies, various features of the presented here analysis tools have been used. First tools have been used for playback and annotation of the activity. Subsequently, the audio and sequences of still images, along with the logfiles of the studies were inserted in the ActivityLens environment through which the goal structures of the activities were constructed and studied.
In (Voyiatzaki, Christakoudis, Margaritis & Avouris, 2004) a study is discussed of activities that took place in a computer lab of a high school, using *Synergo*. The logfiles of Synergo were analysed along with contextual information in the form of video recording of the classroom during the activity and with observers' notes. These where interrelated and through this the verbal interventions of the tutor where identified and the effect of these on the students problem solving process was studied. This study identified the patterns of pupils' reaction to tutoring activity
Finally, in a third case, the collaborative learning activity about a mystery play in a Museum using PDAs, has been studied (Cabrera et al., 2005), (Stoica, Fiotakis, Simarro, Muñoz Frutos & Avouris, 2005). In the study, a logfile of the Museum server was studied in relation to three streams of video from different angles together with the observers' notes. It was found that various events related to interaction of the students with the exhibits and verbal interactions of the students between them and with their tutor/guide were captured in the video streams and were interrelated with actions at the user interface level of the various PDAs that were automatically recorded by the software application used. In this particular study it was found that the additional information conveyed through the posture of the users, their spatial location etc, was

Symposium Epal

important for studying and understanding the activity, while the limited size of the portable devices and the technical limitations of monitoring the PDA screens during the activity, made the video streams and interrelated logged events at the side of the server most valuable source of information.
A summary of the presented and briefly discussed studies is included in Table 1. In the three studies, the common characteristic was that in order to analyse effectively the studied activities and test their hypotheses the analysts used additional evidence in various forms, mostly video and audio. These were added to logfiles generated by the software tools used (chat messages exchanged, actions on concept mapping tools etc.) and were interrelated to them. The analysis environment ActivityLens that was used in these cases facilitated and supported effectively the analysis and evaluation task.

Table 1. Summary of the presented case studies

Study	Setting	Data Sources	Mode of collaboration	Use of ActivityLens
Fidas et al. 2005	Technical Lyceum, Information Technology class, 20 pupils	Logfiles Observer notes audio	*ModelsCreator3* through the network, and face to face	Interrelation of computer based activity and recorded face to face interaction, patterns of collaboration emerged
Voyiatzaki et al. 2004	Junior High School, Computer Lab, 20 Pupils	Logfiles Video Observer notes Activity sheets	*Synergo* through the network, with tutor intervention	The teacher intervention was recorded in video and the effect on students activity was identified
Stoica et al. 2005, Cabrera et al. 2005	Historical/Cultural Museum activity School party (15 year olds) of 12 pupils	Logfiles Video Observer notes	Face to face, Using wireless network-enabled PDAs	Students gestures, posture and face to face interaction captured on video and interrelated to logs of PDAs and screenshots

Symposium EPAL

4. Towards a common format for logging and annotation of interaction data

Given the importance of the field of collaborative learning and the need for standard methods and tools for interaction analysis of recorded data, the need arises for defining a model that describes semantically and syntactically data generated during this process. Kahrimanis et al. (2006) have contributed towards such a model that is directly related to the Synergo and ActivityLens environments discussed in the previous sections. The needs that motivate the development of such a model include the fact that often collaboration support software tools log data in order to fulfil just internal functionality needs regardless of any concerns on interoperability with other tools. A noticeable exception to this is found in cases of tools which may have been developed by the same developers. However, a common logging format would offer new possibilities towards interoperability among distinct tools.
Some classic architectures of Computer Supported Collaborative Work (CSCW) tools Calvary et al. (1997), Dewan (1995) allow sharing low level information that is interpreted by the environment of each user and reproduced according to certain environment's user interface functionalities. In the case of a common logging format, it would be possible that two different tools could be used in the same collaborative session regardless of their additional functionalities for representing log file data in

the user interface. Other prerequisites for the realization of such a scenario is that distinct tools function equivalently at the level of user interface constituent components (widgets) and a common protocol for coordination messages is also shared.

4.1 Compatibility between collaboration tools and analysis tools

Analysis of a collaborative activity constitutes an important part of the integration of collaborative activities in an educational environment. Teachers need to analyze these activities in order to assess their students in terms of grades, to evaluate the activities in order to examine their impact on the wider educational activity and improve their practices according to the evaluation findings and conclusions. Researchers and tool developers also need to analyze data in order to shed light into important aspects of collaborative practices and improve the tools used accordingly.

Many analysis tools have been developed to aid research in the broader scope of behavioral sciences. These tools face the limitation that most data they use as input are not structured and need hand-coding in order to be processed automatically. On the contrary, analysis tools designed specifically for collaboration evaluation studies, like the ActivityLens discussed in section 3, take advantage of data captured automatically by relevant tools. However, it is usually the case that each tool can handle data produced by a limited number of related collaboration support tools

A solution to this problem may be supported by the definition and adoption of a common logging format. In such case, diverse analysis tools providing support for various analysis techniques would be able to manipulate data produced by any collaboration support environment. This widens the horizon of research from various methodological standpoints. It offers possibilities that would not easily be feasible if one had to be restricted to certain tools with specific analysis orientations.

There is also the case that compatibility between collaboration tools and analysis tools is needed both ways. Although this seems paradoxical, it constitutes a real need because in many cases collaboration environments, like Synergo, contain integrated analysis tools and analysis results may need to be fed back to the collaborating partners.

Symposium EPAL

4.2. Desirable properties of the model

In order for the model to cover the external demands stated in the previous section, it should satisfy the following characteristics.

First of all, it has to be ***interoperable.*** It should therefore be formally defined and structured in a standard language. Kahrimanis et al. (2006) developed such a model as an XML Schema since it is a widely accepted standard that is sufficient for modelling purposes.

It should also be *generic* enough so that data of different sources and nature can be described in the model. There is a trade-off between the need for structure that interoperability imposes and the disparity of different data sources. The use of required and optional elements and the abstract form that some elements' meanings selectively take, is one mechanism used for setting an optimal balance for this trade-off.

There is also the need for *flexibility* of the model. Apart from being generic, the model should also be adaptable in various modes of usage according to the needs of different researcher standpoints. Thus, we define a model that is both human and machine interpretable. Moreover, flexibility refers to the need for the model to be applicable in specific circumstances with peculiar needs. For this reason space is provided in the model for some XML elements without a specified meaning, which can include information that is not explicitly described elsewhere in the model.

In order for the model to be actually used, the functionality of potentially compliant tools should be changed, so that they log data in an XML file following the model's format. However, this approach demands a great deal of effort, in terms of resources and time, which is not always affordable to the institutions or companies that are proprietors of the tools.
An alternative approach is that the tools' logging mechanisms remains the same and their log file is transformed into an equivalent log file compliant to the model. In case a tool logs data in the form of an XML file, XSL Transformations (XSLT) technology can be used for the conversion. In cases where a tool stores data in a database or uses another markup language such as RDF, XML parser applications may be used.

4.3. Description of the model

The development of the model followed a spiral process. It started from examples of typical CSCL tools and was followed by the examination of other tools one after the other.

ModellingSpace and *Synergo* were the first CSCL tools taken into account. Afterwards, *ColAT* and *ActivityLens*, which are compliant in a certain extent to the logging format of the aforementioned tools, were also tested. New elements that provided support for extra analysis facilities of ColAT and ActivityLens, such as multiple-level grouping of actions were added to the model. Two non-CSCL open source tools, namely *PHPbb* and *Moodle*, were then tested. The process went on with the study of Noldus *Observer*™, which is a widely used analysis tool, suitable for a wide spectrum of research areas. Finally a similar model developed in the scope of Kaleidoscope (Harrer etal, 2005) was examined. This model aims at providing unidirectional compatibility from CSCL tools to analysis tools. Nine learning support tools (four of which are CSCL tools) and seven analysis support tools were taken into account while developing that model. The authors of the model committed to provide the functionality for transforming the logging mechanism of all these tools so that they are compliant to their model. Our model was tested and was in large extend in concordance with model discussed in (Harrer etal, 2005). The mapping among the models has been defined so that our model can be compliant to all the tools taken into account in (Harrer etal, 2005).

This section presents the main aspects of the model. The first part describes the *context* of the model and the second the *actions* (communicative or gestural) taken place during an activity.

The context part of the model was included due to several reasons: Firstly, there is a certain need for providing a structured description of the context of a CSCL activity. Secondly, data that refer to *user* descriptions, their *roles* and their assignments in *groups* should be stored once and not repeated recurrently accompanying actions. In

Symposium Epal

addition, the description of the tools that are used and the information about the user-interface object types that they provide are included in the context part in favour of *modularity* of the model. Each action logged makes an explicit reference to the identifier of an object type and the property that it causes to change, instead of repeating object type specific information needlessly.

The first part of the action branch of the model concerns modelling of actions that commonly occur in collaborative activities. Concerning this part, what discriminates the model from previous approaches (Martinez et al. 2003) and constitutes its innovative proposal is the way objects are instantiated and referenced. As stated in the previous section, an object is instantiated with a certain action. Its type, the tool that it belongs to and its property types (that an action may cause to change), are referenced in the context part of the model. In addition, a further distinction to previous models is that the model explicitly supports non-collaborative tools, even hand-scripted interactions occurring during face-to-face interactions.

The second part of the action part of the model includes information that relates to data used for analysis of collaborative activities. The most common and trivial action that an analysis tool facilitates and the model supports is the addition of a comment or an annotation to an action. The model takes also into account cases in which an action is classified according to a coding scheme. Usually, messages are coded and assigned to a category of a classification scheme so that they indicate an "agreement", a "proposal" etc. The coding may be defined by the users prior to posting the messages (explicitly: by selecting a category, or implicitly: by choosing sentence openers mapped to categories). Alternatively, the assignment into categories may take place a posteriori during the analysis of the activity by researchers or teachers engaged in content analysis studies.

However, in that way, the model prescribes log files that are just treated as an input from the analysis tools. Thus, we proceeded the development of the model and integrated descriptions of actions taken by researchers and logged by analysis tools.

A more sophisticated facility of analysis tools is the building of an interpretative model of an activity in the form of a multilevel structure, like the one included in ActivityLens. In our model we provide a complex element called *level annotation* that contains the name of the level as an attribute (e.g "Action level") and the name of the certain instance of that level as a sub-element. Consequently, the model satisfies the need of a multilevel interpretation and it not restricted to certain interpretations of that kind as well.

4.4. Validation of the collaborative data model

Log files produced during real-world collaborative activities by all the tools mentioned in the previous section were transformed into their equivalents in the format of the Interoperability Model in order to validate it. In this section we present just an example of such validation. A typical example of a logged action (dispatch of a chat message) in ModellingSpace format is shown in Fig.14 .

Symposium Final

```
<event>
  <time>09 : 47 : 55</time>
  <time2>00 : 07 : 44</time2>
  <user>thodoris</user>
  <action>Chat message</action>
- <attribute>
    [isws kai enas server na einai ontothta giati exei p.x mia dieuthunsh ]
  </attribute>
  <typology/>
  <comments/>
  <added_by_user>false</added_by_user>
  <id_event>49</id_event>
</event>
```

Figure 14 - Action log according to ModellingSpace's format

The equivalent log in the model we presented would take the form that is shown in ***Figure*** .

```
- <action id="49" abs_time="09:47:55" rel_time="00:07:44" successful="1">
    <user_ref>2</user_ref>
    <type>Chat_change</type>
    <object_type_ref>3</object_type_ref>
    <object_ref>41</object_ref>
  - <property_change property_ref="1">
      isws kai enas server na einai ontothta giati exei p.x. mia diauthunsh
    </property_change>
  </action>
```

Figure 15 – Action log according to Interoperability model format

5. Cases of use of Synergo and ActivityLens

In this section, we provide an overview of known research or practice involving the tools and models presented in this document. Numerous research groups have used Synergo and ActivityLens. As a result a corpus of material has been generated in the form of logfiles and analysis files. Some groups are involved in studies in the context of joint projects whereas others have just used these tools developed by the HCI Group in the context of their own research activities.
The most notable uses of the tools that have been reported to the HCI Group have been made by the following groups:

-***Synergo*** (http://hci.ece.upatras.gr/synergo/)

- Collide, University of Duisburg-Essen, Germany
- Department of Psychology, University of Freiburg, Germany
- Universidade Católica de Brasília (UCB), Brasília, Brazil

Modelling Space (Margaritis et al. 2003)

-Departamento de Ingenieria Electrica, Electronica y Control,

Escuela Técnica Superior de Ingenieros Industriales (U.N.E.D), Madrid, Spain
University of the Aegean (GR),
The University of Angers, (F),
The New University of Lisbon and high schools of the Lisbon region (PT)
The University of Mons-Hainaut and schools of the region (B),

ActivityLens (Avouris, Komis, Margaritis & Fiotakis, 2004)

- Department of Educational Sciences and Early Childhood Education, University of Patras, Greece
- LIFC – Université de Franche-Comté, France
- GRIC-COAST, CNRS & Université Lumière Lyon 2, France
- APL group, Computer Science Department, Ecole des Mines, Saint-Etienne, France

In the scope of the Computer-based Analysis and Visualization of Collaborative Learning Activities (CAViCoLA) European Research Team (ERT) (http://cavicola.noe-kaleidoscope.org/) of the Kaleidoscope Network of Excellence (NoE) (http://www.noe-kaleidoscope.org), several studies have been organised that involve research teams using tools developed by the HCI team. Two joint studies have been conducted with the Collaborative Learning Intelligent Distributed Environments (Collide) Group at the University of Duisburg-Essen, Germany that is characterised by a similar research profile. Both studies involved synchronous computer-supported collaboration of dyads consisting of one university student in Greece and one in Germany engaged in collaborative problem solving. Half of the dyads used the Synergo tool and half of them an equivalent tool by developed by the Collide Group, namely Freestyler (Gaßner, 2003) (http://www.collide.info/software/pscproject.2006-09-20.0892820471). The aim of the studies (apart from evaluating learning performance in cross-national settings) was to convey a joint analysis of data gathered by different tools and use the analysis tools of the two teams in an equivalent way. This constitutes the first step towards making these tools more interoperable and managing to eventually use corpora gathered by one tool by another. Results of the second study are reported in (Harrer et. al., 2006). Corpora analysed in these studies involve logfiles of Synergo and Freestyler combined with logs from various multipurpose network tools that were actually used by the students such as instant messaging tools, asynchronous discussion forums and e-mailing tools.
Another case of collaboration of the HCI Group that involves the usage of Synergo by another research team is on progress and takes place also in the scope of CAViCoLA ERT. The study involves the application and further development of a rating scheme for assessing the quality of computer-supported collaboration processes developed by (Spada et. al., 2005, Meier et. al. 2007) to typical CSCL activities conducted with the use of Synergo. The research activity involves the HCI Group and the developers of the scheme from the Department of Psychology at the University of Freiburg, Germany. The pluralism in expertise by the two teams and backgrounds of researchers is considered to be beneficial for the research activity and possibly reveal new directions of further innovation and development of the tools. The first version of the rating scheme was developed based on evaluation of learning activities that involved dyads of university students with prior discrepancies in knowledge communicating through teleconferencing systems. The application and refinement of the model by its application to studies with Synergo is a process that is still evolving and involves the collaboration of HCI Group with the research team of the

Symposium Epal

Department of Psychology, Albert-Ludwigs-Universitaet, Freiburg,, Germany. Apart from the use of Synergo as a collaboration support tool in these studies, its' integrated analysis tools and especially the facility to reproduce the activity described below in this paper are used for evaluation purposes. Corpora, used in this case, mostly constitute of Synergo logfiles,

Another study, in the scope of which ActivityLens is being used, is currently in its intitial phase. It involves students collaborating closely using cognitive tutors (Anderson, Corbett, Koedinger, & Pelletier, 1995) in the domain of algebra. The studies were organised and conducted by the research team of the University o Freiburg mentioned in the previous paragraph and the Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, USA. ActivityLens is going to be used as a tool that facilitates integrated handling of logfiles of tutors and videocaptures (apart from other facilities of the tool that are also considered to be useful). The activities under inspection follow experimental methodology design and include both individual students interacting with tutors and collaborative sessions as experimental conditions. In the current phase of the study, researchers work at making logfiles produced by cognitive tutors compatible to ActivityLens .

Another potential use of ActivityLens deals with its use for analysis of argumentation diagram building activities conveyed with the use of tools such as DREW (Corbel et. al., 2002). This can provide an opportunity to use the tool for analysis of CSCL that involve different tasks than the typical ones analysed by the HCI Group. The application of the methodology of the rainbow framework (Baker et. al. 2003) can possibly inspire the design of new facilities of the tool or reveal shortcomings and new directions for existent facilities. On the other hand, researchers in Université Lumière Lyon 2, France, - APL group, and the Computer Science Department, Ecole des Mines, Saint-Etienne are interested in analysing studies using the ActivityLens tool and getting ideas on how to proceed with the implementation of a similar analysis tool they are currently developing. The corpora that relate to these studies are currently variant and follow a format that is tightly related to the logging tool. Corpora used by ActivityLens follow the format of the HCI Group that is also common for Synergo and tools like DREW log files in a format that is not currently compatible. Moreover, studies in which the rainbow framework is used usually involve the transcription of oral dialogues which has to be compatible with ActivityLens in order to analyse such data with the tool.

Synergo and ActivityLens have also been used independently by other research teams such as Universidade Católica de Brasília (UCB), Brasília, Brazil, and LIFC – Université de Franche-Comté respectively.

Symposium Epal

Conclusion

There is a dire need for a common description of corpora captured in learning activities. For the field of collaborative learning this can by justified by various standpoints.

First, analysis of a CSCL activity constitutes an important part of the integration of CSCL activities in an educational environment. Teachers need to analyze CSCL activities in order to assess their students in terms of grades, to evaluate the activities in order to examine their impact on the wider educational activity and improve their practices according to the evaluation findings and conclusions. Researchers and tool developers also need to analyze data in order to shed light into important aspects of

CSCL practices and improve CSCL tools accordingly. Many analysis tools have been developed to aid research in the broader scope of behavioural sciences (e.g. Noldus, 1991). These tools face the limitation that most data they use as input are not structured and need hand-coding in order to be processed automatically. On the contrary, analysis tools designed specifically for CSCL evaluation studies (Avouris et. al. 2004) take advantage of data captured automatically by CSCL tools. However, it is usually the case that each tool can handle data produced by a limited number of related CSCL tools. A solution to this problem may be supported by the definition and adoption of a common logging format. In that case, diverse analysis tools providing support for various analysis techniques would be able to manipulate data produced by any CSCL tool. This widens the horizon of research from various methodological standpoints. It offers possibilities that would not easily be feasible if one had to be restricted to certain tools with specific analysis orientations. There is also the case that compatibility between CSCL tools and analysis tools is needed both ways. Although this seems paradoxical, it constitutes a real need because in many cases CSCL environments contain integrated analysis tools and analysis results may need to be fed back to the collaborating partners.

Another crucial need in the field of CSCL analysis that has emerged during the late years of extended diffusion of common purpose network tools and is often underestimated has to be emphasised. Researchers and activity designers in the field of CSCL sometimes develop strict educational scripts, provide certain CSCL tools to the students and restrict the students to conduct an activity according to their directives (Dillenbourg, 2002). However, in practice, students prove to be surprisingly flexible in terms of computer tool usage: they adopt alternative media in order to interact with their peers. Usage of mailing applications, instant massagers and asynchronous discussion forums are the most common examples. This reality introduces a new problem when analyzing CSCL activities: Researchers examine CSCL interaction using the logs of the tools that they provide, missing important information conveyed via other channels of communication. This introduces the need for a *holistic description of data* used as input for analysis, i.e. including all computer tools used during the CSCL activity regardless whether they are CSCL-specific or not. Furthermore, in the case of face-to-face CSCL activities, other data, like video or even hand-coded data that describe other channels of communication such as oral dialogue or gestures, should be integrated in a common description.

In order to satisfy such needs, members of the HCI Group are working towards the direction of a common description of corpora of learning activities. First of all, HCI Group members have proposed a model for interoperability in Computer Supported Collaborative Learning (Kahrimanis et. al., 2006). Moreover, the CAVICOLA ERT of Kaleidoscope Network of Excellence mentioned above involves a work package on Tool Interoperability that is tightly related to that goal. Part of this task is the adoption of the research teams involved of a common way to log and describe learning data. Moreover, this approach is also shared by another project that some teams in Cavicola are also members of, namely the ARGUNAUT project (http://www.argunaut.org/).

When such initiatives mature enough, the benefits of the research teams mentioned above to exchange their corpora would be multifaceted in such a recent research field as computer supported collaborative learning.

Capitalization of findings of studies is something that has not gone too far in the area of CSCL. Meta-analyses revealing interesting aspects of CSCL activities that are not conceivable by examining isolated studies may come up. Research groups working on similar learning cases may benefit from each other's analyses and build on them. A

Symposium EPAL

straightforward example is the one of collaboration between the HCI Group and University of Duisburg-Essen mentioned above. Since, a large part of the studies conducted by these two teams deal with similar cases, and therefore datasets for analysis, it would be valuable to have these data on a common repository and manipulate them according to one's wishes. Large corpora could be also valuable for extended statistical analyses of data occurred form real-world learning activities which is difficult with current settings.

During the process of developing the Interoperability Model, discussed in section 4, it was found that most log files of collaboration tools use equivalent semantics to store and describe log file data despite their technical variations. This implies that no major modifications would be needed in order for most tools to comply with the interoperability model.

Concerning analysis tools, we noticed that they provide flexible formats that can be appropriate for various analysis approaches, however not much attention is paid to the output that they produce. This imposes a problem for such tools to be fully compatible to the interoperability model. Additional functionality should be supported so that analysis tools can create log files according to the model.

The mapping of non-CSCL tools was also proved to be semantically straightforward, but some of these tools also have the problem of logging data just for their internal functionality purposes (usually in a database), regardless of interoperability concerns. However, we believe that it is worth the effort to make tools used in a CSCL activity compliant to the interoperability model, due to the new range of possibilities that this standard would provide.

Further work, extending this research thread can follow two directions. Firstly, concerning the development of the model, more tools of wider diversity could be used in order to test and reform the model. We noticed that the model converged to a certain form as its formative evaluation in relation to certain tools proceeded. However, there is always space for further reform or enrichment of the model with additional features. Secondly, further long-term application of the model in various real-world circumstances should reveal possible shortcomings of the model or inspire further research. The direction of proposing the model as a standard and making it compatible with other existing standards in the field may enhance its use and should be pursued.

The development of new analysis tools or the improvement of existing ones could also potentially benefit from the seamless availability of rigorously described corpora of CSCL data. Analysis tools that have gone through testing against certain kinds of data produced by certain kinds of studies can be further tested and refined using richer and more diverse data as an input. For example, the design of a future extended version of the ActivityLens tool could benefit from its pilot usage to data of other research teams that were collected from other studies and are usually analysed from other methodological points of view.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4, 167-207.

Symposium Epal

Avouris N.M., Dimitracopoulou A., & Komis V. (2003). On analysis of collaborative problem solving: An object-oriented approach, *Journal of Human Behavior*, 19 (2), 147-167.

Avouris N., Komis V., Fiotakis G., Dimitracopoulou A., and Margaritis M.(2004a) Method and Tools for analysis of collaborative problem-solving activities. *Proceedings of ATIT2004*, First International Workshop on Activity Theory Based Practical Methods for IT Design , Copenhagen, Denmark, September 2004, pp. 5-16, available from http://hci.ece.upatras.gr

Avouris N., Komis V., Margaritis M., & Fiotakis G.(2004b). An environment for studying collaborative learning activities. *Journal of International Forum of Educational Technology & Society*, 7 (2), 34-41.

Avouris N., Margaritis M., and Komis V., (2004c). Modelling interaction during small-group synchronous problem-solving activities: The Synergo approach, 2nd International Workshop on Designing Computational Models of Collaborative Learning Interaction, ITS2004, 7th Conference on Intelligent Tutoring Systems, pp. 13-18, Maceio, Brasil, September 2004.

Baker, M. J., Quignard, M., Lund, K., & Séjourné, A. (2003). Computer-supported collaborative learning in the space of debate. In B. Wasson, S. Ludvigsen, & U. Hoppe (Eds.), *Designing for change in networked learning environments* (pp. 11–20). Dordrecht: Kluwer Academic Publishers.

Bertelsen O.W., & Bodker S. (2003), Activity Theory. In J. M Carroll (Ed.), *HCI Models, Theories and Frameworks*, San Francisco, CA, USA: Morgan Kaufmann.

Calvary, G., Coutaz, J., and Nigay, L: From Single-User Architectural Design to PAC*: a Generic Software Architecture Model for CSCW, Proceedings of CHI 97, ACM publ., 1997, pp. 242-249.

Dewan, P, "Multiuser Architectures", in Proceedings EHCI'95, Working Conference on Engineering Computer Human Interaction, 1995.

Cabrera, J. S., Frutos, H. M., Stoica, A. G., Avouris, N., Dimitriadis, Y., Fiotakis, G., & Liveri, K. D. (2005). Mystery in the museum: collaborative learning activities using handheld devices. *Proc. 7th Int. Conf. on Human Computer Interaction with Mobile Devices & Services*, Salzburg, Austria, September 19 - 22, 2005, vol. 111. ACM Press, New York, NY, 315-318

Corbel, A., Girardot, J.J., Jaillon, P. (2002). DREW: « A Dialogical Reasoning Web Tool » ICTE2002, Int. Conf. on ICT's in Education. Badajoz, Espagne, 13-16 novembre 2002.

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. In P. A. Kirschner (Ed.) *Three worlds of CSCL. Can we support CSCL* (pp. 61-91), Heerlen, Open Universiteit Nederland.

Symposium EPAL

Dix A., Finlay J., Abowd G, Beale R., (1998), Human-Computer Interaction, Prentice Hall.

Edwards, D. & Potter, J. (1992) Discursive Psychology. London: Sage.

Fidas, C., Komis, V., Avouris, N.M., Dimitracopoulou, A. (2002). Collaborative Problem Solving using an Open Modeling Environment, G. Stahl (ed), ***Proc. CSCL 2002,*** pp. 654-656, Boulder Colorado, January 2002, Erlbaum Assoc. Hillsdale, NJ, 2002

Fidas C., Komis V., Tzanavaris S., Avouris N., (2005). Heterogeneity of learning material in synchronous computer-supported collaborative modelling, ***Computers & Education***, 44 (2), 135-154, February 2005.

Garfinkel, H. (1967) Studies in Ethnomethodology, Englewood Cliffs, NJ: Prentice-Hall.

Gaßner, K (2003). Diskussionen als Szenario zur Ko-Konstruktion von Wissen. Dissertation. Faculty of Engineering Sciences, University Duisburg-Essen.

Hammersley, M. (1982). The Sociology of the Classroom. In A. Hartnett (Ed.), ***The Social Sciences in Educational Studies***. London, UK: Heinemann.

Harrer A. et. al.(2005), The Common Format, Chapter 4, in A. Martinez, Library of Interaction Analysis Methods, ***Deliverable D31.2.1, Kaleidoscope NoE*** (available from www.noe-kaleidoscope.org), 2005.

Harrer, A., Kahrimanis, G., Zeini, S., Bollen, L., Avouris, N. (2006) Is there a way to e-Bologna? Cross-National Collaborative Activities in University Courses, ***Proceedings 1st European Conference on Technology Enhanced Learning EC-TEL***, Crete, October 1-4, 2006, Lecture Notes in Computer Science, vol. 4227/2006, pp. 140-154, Springer Berlin

Heath, C. (1986) Video analysis: Interactional coordination in movement and speech. ***Body Movement and Speech in Medical Interaction***, Cambridge University Press, Cambridge, UK, 1-24

Jordan, B. & Henderson, A. (1995) Interaction analysis: Foundations and practice, ***Journal of the Learning Sciences***, 4 (1), 39-103

Kahrimanis, G., Papasalouros, A., Avouris, N., Retalis, S. (2006). A Model for Interoperability in Computer Supported Collaborative Learning. ***Proc. ICALT 2006. The 6th IEEE International Conference on Advanced Learning Technologies***, IEEE Publ., July 5-7 2006, Kerkrade, Netherlands, 51-55.

Komis V., Avouris N., Fidas C. (2002), Computer-Supported Collaborative Concept Mapping: Study of Synchronous Peer Interaction, ***Education and Information Technologies***, 7:2, 169–188

Margaritis M., Avouris N., Komis V., (2004), Methods and Tools for representation of Collaborative Learning activities. Proc. ETPE 2004, September 2004, Athens.

Symposium Epal

Martinez A., de la Fuente P., and Dimitriadis Y.,(2003) , "An XML-based representation of collaborative interactions", In B.Wasson, S. Ludvigsen & U. Hoppe (Eds) Computer Support for Collaborative Learning: *Designing for Change in Networked Learning Environments,* (CSCL 2003), Bergen, Norway, 2003, pp. 379-384

Meier, A., Spada, H., & Rummel, N. (2007). A rating scheme for assessing the quality of computer-supported collaboration processes. To appear at: *The International Journal of Computer-Supported Collaborative Learning*.

Noldus LPJJ, (1991). The Observer: a software system for collection and analysis of observational data. *Behav. Res. Methods Instrum* 2, 1991, pp. 415–429

Noras, M. (2006). Un besoin de spécifications des corpus de formation en ligne. 1ères Rencontres Jeunes Chercheurs en EIAH, RJC-EIAH'2006, pages 165 à 172

Spada, H., Meier, A., Rummel, N. & Hauser, S. (2005). A new method to assess the quality of collaborative process in CSCL. In D. Suthers & T. Koschmann (Eds.), *The next 10 years! Proceedings of the International Conference on Computer Support for Collaborative Learning 2005.* New York, Springer.

Stoica A., Fiotakis G., Simarro Cabrera J., Muñoz Frutos H., Avouris N., & Dimitriadis Y. (2005). Usability evaluation of handheld devices: A case study for a museum application, *Proceedings PCI 2005*, Volos, Greece, November 2005.

Suthers D. (2001), Architectures for Computer Supported Collaborative Learning. In *proceedings of the IEEE International Conf. on Advanced Learning Technologies* (ICALT2001), 6-8- Aug. 2001. Madison, Wisconsin.

Voyiatzaki E., Christakoudis Ch., Margaritis M., Avouris N. (2004). Teaching Algorithms in Secondary Education: A Collaborative Approach, *Proceedings ED Media 2004, AACE Publ*, Lugano, June 2004, 2781-2789.

Winograd T., (1987). A Language/Action Perspective on the Design of Cooperative Work, *Human-Computer in Interaction* 3:1 (1987-88), 3-30.

Symposium Epal

Référence de publication de cet article : S. Salmon-Alt, L. Romary, J.-M. Pierrel. “Un modèle générique d'organisation de corpus en ligne : application à la FreeBank", Traitement Automatique des Langues, Vol.45, n°3, pp. 145-169, 2004.

Un modèle générique d’organisation de corpus en ligne : application à la *FReeBank*

S. Salmon-Alt* — L. Romary — J.-M. Pierrel***

**ATILF – UMR 7118*
Analyse et Traitement Informatique de la Langue Française
44, avenue de la Libération, BP 30687, F-54063 Nancy Cedex
{salt@atilf.fr, Jean-Marie.Pierrel@atilf.fr}

*** LORIA – UMR 7503*
Laboratoire Lorrain de Recherche en Informatique et ses Applications
Campus scientifique, BP 239, F-54506 Vandœuvre-lès-Nancy Cedex
Laurent.Romary@loria.fr

Copyright Epal

RÉSUMÉ. *Les corpus français librement accessibles et annotés linguistiquement sont insuffisants à la fois quantitativement et qualitativement. Partant de ce constat, la FREEBANK se veut une base de corpus du français annotés à plusieurs niveaux (structurel, morphologique, syntaxique, coréférentiel) et à différents degrés de finesse linguistique qui soit libre d’accès, codée selon des schémas normalisés, intégrant des ressources existantes et ouverte à l’enrichissement progressif. Préalablement à la présentation du prototype qui a été réalisé, le présent article propose une modélisation générique de l’organisation et du déploiement d’une archive de corpus linguistiques dans la continuité des travaux menés au niveau international sur la représentation des ressources linguistiques (TEI et ISO/TC 37/SC 4).*

ABSTRACT. *The few available French resources for evaluating linguistic models or algorithms on other linguistic levels than morpho-syntax are either insufficient from quantitative as well as qualitative point of view or not freely accessible. Based on this fact, the FREEBANK project intends to create French corpora constructed using manually revised output from a hybrid Constraint Grammar parser and annotated on several linguistic levels (structure, morpho-syntax, syntax, coreference), with the objective to make them available on-line for research purposes. Therefore, we will focus on using standard annotation schemes, integration of existing resources and maintenance allowing for continuous enrichment of the annotations. Prior to the actual presentation of the prototype that has been implemented, this paper describes a generic model for the organization and deployment of a linguistic resource archive, in compliance with the various works currently conducted within international standardization initiatives (TEI and ISO/TC 37/SC 4).*

MOTS-CLÉS : *ressources linguistiques, annotation multi-niveau, normalisation, ressources libres…*

KEYWORDS: *linguistic resources, multi-level annotation, standardization, open resources…*

Nom de la revue. Volume X – n° X/2001, pages 1 à X

1. Introduction

L'idée de la *FReeBank* – en tant qu'espace de dépôt, de maintenance, de distribution et de normalisation de ressources libres pour l'étude et le traitement du Français – fait partie des résultats théoriques et pratiques majeurs d'un projet précédent (*Ananas*[1] ; Salmon-Alt, 2002). Ce projet, centré initialement sur l'annotation sémantique de corpus existants, s'est en effet rapidement heurté à l'indisponibilité de corpus pré-annotés, réutilisables et libres de droit. Au-delà d'un important travail de collecte et d'annotation de corpus libres (balisage TEI (Sperberg-McQueen et Burnard 2002), segmentation, étiquetage morpho-syntaxique, analyse syntaxique et annotation anaphorique), cela nous a amenés à réfléchir d'une façon plus générale à une architecture de gestion de ressources linguistiques en ligne. En plus de notre volonté d'encourager le partage de corpus annotés dans la communauté scientifique, l'initiative de la *FReeBank* a été motivée par trois constats positifs : le déploiement de plusieurs campagnes d'annotation, manuelle ou automatique, au-delà du niveau morpho-syntaxique (projets Evalda Easy/Média[2]), d'autres initiatives de mise en ligne ou de recensement de corpus libres (Asila[3], ABU[4], BDCOIFA[5]) et l'avancement des fondements théoriques de la normalisation de ressources linguistiques (Bird et Liberman, 2001 ; Ide et Romary, 2004b et les travaux de l'ISO TC 37/SC 4[6]). Dans ce contexte, la *FReeBank* se veut un espace ouvert de gestion de ressources libres, permettant le dépôt et le téléchargement de données brutes ou annotées, mais aussi le dépôt d'annotations sur des ressources existantes, par exemple sous forme de méta-annotations (validation, dépréciation, correction ou affinage d'annotations précédentes) ou d'annotations concurrentes (annotations multi-annotateur). Si elle n'impose aucune validation *a priori* des annotations soumises, elle met l'accent sur une documentation exhaustive et standardisée des données répertoriées. Cela concerne d'une part la documentation des informations linguistiques apportées par l'annotation et, d'autre part, la documentation des méta-données, en particulier la description de la ressource, des niveaux de description et des contributeurs. Dans cet esprit, la *FReeBank* est conçue comme un espace expérimental ouvert, évolutif et générique, fondé sur une méthodologie de modélisation d'annotations linguistiques qui tente d'allier une analyse fine des pratiques et besoins linguistiques aux initiatives internationales de représentation de données langagières.

Symposium EPAL

[1] http://www.atilf.fr/ananas
[2] http://www.technolangue.net/
[3] http://www.loria.fr/projets/asila/
[4] http://abu.cnam.fr/
[5] http://www.unicaen.fr/corpus/
[6] http://www.tc37sc4.org

2. Fondements théoriques

2.1. *Organisation générique d'une archive linguistique*

2.1.1. La notion de « corpus »

La *FreeBank* repose sur l'idée fondamentale que toute donnée linguistique peut se caractériser par une certaine couverture linguistique, c'est-à-dire un contenu langagier linéaire fini identifié dans un contexte de production particulier. Le contexte de production peut être spécifié de multiples façons. Il peut s'agir par exemple, pour l'écrit, d'une édition particulière d'une œuvre, ou, pour l'oral, de l'identification d'un (ou de plusieurs) locuteur(s) et d'un instant de locution. À partir de cette notion, la *FreeBank* définit un corpus comme une collection de données relatives à une certaine couverture linguistique, vue en tant qu'objet d'études linguistiques ou littéraires. Ce concept couvre, par exemple, des sources audio d'une conversation enregistrée, des textes écrits bruts, des dialogues transcrits ou encore du matériau linguistique existant sous forme de manuscrits anciens numérisés. Dans cette acception, le texte intégral du *Père Goriot* de Balzac, dans l'édition de Gallimard 1976, constitue un corpus différent de celui obtenu par la sélection du chapitre I de la même édition de cet ouvrage. À l'opposé, le texte du chapitre 1 du *Père Goriot* étiquetée morpho-syntaxiquement (Figure 5) et ce même texte annoté en anaphores (Figure 8) seront considérés comme relatifs à un même corpus. Notre notion de corpus est volontairement générale par rapport à certaines positions dans le domaine (McEnery et Wilson 1996, Habert et al. 1997, Véronis 2000). Parmi les critères proposés par McEnery et Wilson (1996), elle reprend celui de la « taille finie » et celui de la « disponibilité sous forme électronique ». En contrepartie, s'appuyant sur des arguments développés en particulier par Kilgarriff et Grefenstette (2003) à propos de la difficulté d'évaluation de la « représentativité » d'un corpus, elle couvre des compilations linguistiques intentionnelles aussi bien que contingentes.

2.1.2. La difficulté de définir l' « annotation linguistique »

La valeur ajoutée d'un corpus linguistique augmente avec le nombre et la qualité des annotations. D'une façon générale, l'annotation consiste à expliciter des informations linguistiques jusqu'alors implicites dans le matériau, en y ajoutant des données méta-linguistiques (Bird et Liberman, 2001 ; McEnery et Wilson, 1996). McEnery et Wilson (1996) distinguent sept types d'annotations linguistiques : l'étiquetage morpho-syntaxique (Figures 4 et 5), la lemmatisation (Figures 4 et 5), l'annotation syntaxique (Figures 6 et 7), l'annotation sémantique (Figure 7), l'annotation discursive (Figure 8) et la transcription phonétique. Si cette classification va fortement dans le sens du postulat selon lequel une annotation linguistique est toujours le résultat d'un processus d'interprétation (Leech 1993), elle soulève aussi certains problèmes. En particulier, elle exclut certains types d'ajouts d'informations, comme la segmentation en chaînes de caractères (Figure 3),

Symposium EPAL

l'explicitation de la structure d'un texte à la TEI (Figure 2) ou l'insertion de bornes temporelles dans une transcription. Si l'on peut en effet considérer que ces cas ne relèvent pas de l'interprétation – notamment parce qu'il n'y a pas de possibilité de divergence entre annotateurs – ils sont toutefois souvent considérés comme annotations, simplement parce qu'il y a explicitation d'informations préalablement implicites (Bird et Liberman, 2001). Plus généralement, comme le font remarquer à juste titre Leech (1997) et Véronis (2000), il peut être difficile de départager ce qui relève de la représentation de ce qui relève de l'interprétation. Ceci est particulièrement vrai pour les transitions entre « recueil de données » et « transcription », où la part de l'interprétation peut être importante, par exemple lors de la transcription d'enregistrement oraux ou lors du balisage de divisions hiérarchisées dans un texte. Enfin, il peut très bien exister toute une série de représentations successives avant même d'arriver à une annotation au sens classique (fichier audio, transcription phonétique, transcription orthographique, balisage structurel de base à la TEI, segmentation, annotation morpho-syntaxique). Vouloir, dans ces cas, maintenir une séparation entre « représentations » et « annotations » poserait non seulement la question du choix du « matériau primaire » (c'est-à-dire celui qui est supposé servir de référence objective vis-à-vis de la couverture linguistique et donc de toute autre activité d'annotation), mais aussi celle de la justification du statut particulier attribué à celui-ci. Ainsi, si l'on tient vraiment à identifier ce matériau primaire, il devient difficile de choisir entre une stratégie qui sélectionnerait un ensemble hétérogène de niveaux les plus « bruts » parmi ceux disponibles (fichier audio, transcription orthographique, et pourquoi pas annotation morpho-syntaxique...), au risque d'y introduire des données subjectives, ou le choix forcé d'un niveau particulier de représentation, au risque que celui-ci ne soit pas toujours présent dans un corpus donné (par exemple lorsque l'on ne dispose que de la transcription dans le cas de données orales).

Symposium EPAL

2.1.3. De l' « annotation » aux « niveaux de description »

Partant de ces constats – et en radicalisant en quelque sorte encore la position adoptée par Véronis (2000) ainsi que par Bird et Liberman (2001) – nous proposons d'unifier les notions de « recueil de données », « transcription » et « annotation » par l'introduction de la notion unifiée de « niveau de description ». Nous entendons par là tout ensemble cohérent d'informations explicites relatif à un corpus (au sens de notre définition). Par rapport à la notion d'annotation discutée ci-dessus, nous maintenons donc le critère d'explicitation d'information, mais nous n'imposons plus de contraintes sur la nature de ces informations (elles peuvent être linguistiques, mais aussi temporelles ou structurelles), ni sur la part de l'interprétation dans ce processus. Dans cette acception, un niveau de description, toujours relatif à un corpus donné, couvre tout aussi bien un enregistrement audio, du texte brut ou formaté (Figure 1), du texte balisé structurellement (Figure 2), du texte segmenté (Figure 3) et des « annotations » au sens classique, quelqu'en soit le format (Figure 4 à 8). D'un point de vue plus pratique, la question de l'existence d'un « matériau brut » est reposée dans un contexte différent : plutôt que de nous interroger sur les

critères informationnels permettant d'attribuer un statut privilégié à tel ou tel ensemble de données, nous nous interrogeons sur la dépendance entre niveaux de description.

Pour la *FReeBank*, nous avons ainsi identifié que l'une des informations essentielles qui devait être attachée à un niveau de description donné (ou un groupement de niveaux de description dans le cas d'annotations internes, cf. *supra*), est son lien de dépendance éventuel vis-à-vis d'un autre niveau de description. Ce lien établit que le niveau de description doit être complété par les informations d'autres niveaux (éventuellement par transitivité) pour être parfaitement exploitable. Ainsi, on pourra marquer la dépendance entre une transcription orthographique et le fichier audio qui lui a servi de source, ou encore entre une annotation référentielle basée sur une annotation syntaxique identifiant les groupes nominaux (Salmon-Alt et Romary, 2005).

Combiné avec la notion de couverture linguistique, la dépendance entre deux niveaux de description permet maintenant de fournir un cadre plus rigoureux à la définition de ce que l'on peut identifier, au sein d'un corpus donné, comme étant du matériau primaire ou secondaire, et ce par le biais d'une propriété caractérisant spécifiquement un niveau de description. Un niveau de description est ainsi considéré comme « secondaire », dès lors qu'il est nécessaire de faire référence à un autre niveau de description avec lequel il est en dépendance pour en reconstituer la couverture linguistique. Ce sera par exemple le cas d'un étiquetage morpho-syntaxique (FIgure 5) qui pointerait, sans la dupliquer, sur une transcription phonétique ou une segmentation en unité de référence (Figure 3). La reconstitution de la couverture peut se faire par transitivité, l'essentiel étant un continuum de dépendance d'un niveau de description secondaire vers, à un moment donné, un niveau primaire. Tout niveau qui n'est pas strictement dépendant d'un autre pour la reconstitution de la couverture linguistique sera alors considéré comme « primaire ». Il faut toutefois noter que cette notion dépend *in fine* de ce que l'on considère comme la couverture linguistique. Par exemple, une transcription d'un enregistrement audio, alignée par des bornes temporelles avec l'information sonore (cf. recommandations de la TEI), permet de reconstituer partiellement la couverture linguistique du corpus (d'un point de vue orthographique), tout en comportant des pointeurs sur un autre niveau de description qui complète la connaissance que l'on peut avoir de cette même couverture linguistique. Dans ce cas, les deux niveaux de description sont considérés comme étant primaires. Il peut en être de même pour une annotation référentielle dans laquelle des segments linguistiques (les expressions référentielles) sont reliés à des représentations identifiantes pour des objets du contexte (Anderson et al., 1991 ; Salmon-Alt, 2001).

Par ailleurs, on ne peut introduire les notions de niveau de description et de dépendance sans préciser les conditions de représentation effective d'un ensemble de niveaux de descriptions au sein d'un même corpus : un niveau de description donné peut ainsi être physiquement intégré ou superposé à un autre niveau de description ou représenté de façon indépendante. Dans ce dernier cas, s'il y a des

Symposium EPAL

liens de dépendance vers un autre niveau de description, il faudra envisager des mécanismes de référenciation (pointeurs). Ces deux modes de représentation correspondent aux deux notions d' « annotation interne » (*inline markup*) et d' « annotation externe » (*stand-off markup*), introduites au milieu des années 90 par plusieurs auteurs (Ide et Véronis, 1995; Ide et Priest-Dorman, 1996; Thomson et McKelvie, 1997). Une annotation interne est une représentation simultanée de plusieurs niveaux de description au sein d'un même objet informationnel : par exemple, structure, sémantique et coréférence (Figure 2), morpho-syntaxe et syntaxe (Figure 6) ou syntaxe et sémantique (Figure 7). Elle introduit de fait des relations d'ordre et de hiérarchisation entre éléments informationnels appartenant à différents niveaux de description. Or, ces relations peuvent être intentionnelles, mais aussi contingentes : dans un court dialogue de renseignements, il se pourrait, par exemple, que les tours de parole correspondent systématiquement à des actes dialogiques, sans que cela puisse être considéré comme une relation générale et systématique. À l'opposé, l'annotation externe sépare physiquement les niveaux de description et explicite de ce fait les relations de dépendance éventuelles entre ceux-ci par des mécanismes de référence, par exemple des pointeurs (Figures 3 et 5). Pour l'exemple dialogique précédent, l'introduction d'une dépendance explicite entre tours de parole et actes de dialogue est alors peu probable et irait à l'encontre du bon sens linguistique : cela empêcherait par exemple de tenir compte d'un acte dialogique réparti sur deux tours de parole ou d'un tour de parole associé, simultanément ou non, à plusieurs actes dialogiques. En contrepartie, introduire une dépendance entre les niveaux de description morpho-syntaxique et syntaxique semble approprié, puisque la délimitation des constituants syntaxiques repose fondamentalement sur la segmentation opérée au niveau morpho-syntaxique. En pratique, cela se traduira alors pour un pointage de l'annotation vers les unités de référence issues de l'annotation morphologique.

Parmi ces deux modes de représentation, l'annotation externe est le mode le plus générique, généralement recommandé pour une meilleure gestion d'un corpus (Mengel et al., 2000 ; Ide et Romary, 2004b). En raison de l'explicitation des dépendances entre niveaux de description, il permet en effet la représentation parallèle d'un nombre arbitraire de niveaux de description, éventuellement non hiérarchiques, autorise la co-existence de plusieurs versions concurrentes d'un même niveau de description (section 3.2.1.) et permet de modifier, voire supprimer des informations sur un niveau de description particulier sans rendre inutilisables les autres informations. Dans le contexte de la *FReeBank*, où l'intention est justement de ne pas imposer, *a priori*, un modèle particulier d'annotation à toute la communauté, nous donnons la possibilité de déposer des données reposant sur l'un ou l'autre modèle, même au sein d'un même corpus.

2.1.4. L'unité de dépôt physique : la « ressource »

Enfin, l'organisation conceptuelle (ou linguistique) d'un corpus en niveaux de description se superpose avec son organisation physique. D'un point de vue physique, un corpus déposé à la *FreeBank* est organisé en *ressources*. La ressource

Symposium EPAL

est l'unité de dépôt, c'est-à-dire le fichier électronique soumis par un dépositaire. Il convient de noter qu'il n'y a pas d'isomorphisme entre l'organisation conceptuelle et l'organisation physique d'un corpus. Un même corpus peut en effet se présenter sous forme d'une seule ressource, comportant simultanément plusieurs niveaux de description : c'est par exemple le cas de la sortie de l'analyseur syntaxique *VISL-FraG*[7] (Bick, 2003), comportant des informations strictement syntaxiques (constituants et dépendances), mais aussi les résultats d'un étiquetage morpho-syntaxique et d'une lemmatisation (Figure 6). A l'opposé, les informations relatives à un même niveau d'annotation peuvent être « éparpillées » sur plusieurs ressources : c'est par exemple le cas de la sortie du résolveur d'anaphores *Art Nouveau* (Vieira et al., 2003), séparant en unités de dépôts distinctes les listes des expressions anaphoriques, des antécédents et des liens anaphoriques.

Texte brut

Madame Vauquer, née De Conflans , est une vieille femme qui, depuis quarante ans, tient à Paris une pension bourgeoise établie rue Neuve-Sainte-Geneviève , entre le quartier latin et le faubourg Saint-Marceau. Cette pension, connue sous le nom de la Maison-Vauquer , admet également des hommes et des femmes, des jeunes gens et des vieillards, sans que jamais la médisance ait attaqué les mœurs de ce respectable établissement.

Figure 1. *« Goriot » (extrait) : texte source*

Structure, entités nommées et anaphores

<seg><rs type="person-oeuvre" id="p1"><name type="person-oeuvre" key="Mme Vauquer">Madame Vauquer</name>, née<name type="person-oeuvre" key="De Conflans">De Conflans</name></rs>, est une vieille femme qui, depuis quarante ans, tient à <rs type="place-ville" id="pl1"><name type="place-ville" key="Paris">Paris</name></rs> <rs type="org-oeuvre" id="or1">une pension bourgeoise établie<rs type="place-rue" id="pl2"><name type="place-rue" key="Neuve-Sainte-Geneviève">rue Neuve-Sainte-Geneviève</name></rs>, entre <rs type="place-quartier" id="pl3">le<name type="place-quartier" key="latin">quartier latin</name></rs> et le <rs type="place-rue" id="pl4"><name type="place-rue" key="Saint-Marceau">faubourg Saint-Marceau</name></rs></rs>.</seg><seg><rs type="org-oeuvre" id="or2">Cette pension, connue sous le nom de la<name type="org-oeuvre" key="Maison-Vauquer">Maison-Vauquer</name></rs>, admet également des hommes et des femmes, des jeunes gens et des vieillards, sans que jamais la médisance ait attaqué les moeurs de <rs type="org-oeuvre" id="or3">ce respectable établissement</rs>.</seg>...

Figure 2. *« Goriot » (extrait) : TEI structure et sémantique (Bruneseaux et al. 1997)*

[7] http://visl.sdu.dk/visl/fr/parsing/automatic/

Segmentation

```
<word id="word_27">Madame</word>
<word id="word_28">Vauquer</word>
<word id="word_29">,</word>
<word id="word_30">née</word>
<word id="word_31">De</word>
...
```

Figure 3. *« Goriot » (extrait) : texte segmenté*

Morpho-syntaxe 1

1	Madame	madame	NCFIN	Ncf.
2	Vauquer	Vauquer	NPI	Np..
3	,	,	PCTFAIB	Ypw
4	née	naître	VPARPFS	Vmpasf
5	De	de	PREP	Sp
6	Conflans	Conflans	NPI	Np..
7	,	,	PCTFAIB	Ypw
8	est	être	VINDP3S	Vmip3s
9	une	un	DETIFS	Da-fs-i
...				

Figure 4. *« Goriot » (extrait) : analyse morpho-syntaxique (Cordial)*

Morpho-syntaxe 2

```
<w span="word_27"   msd="SBC:_:s"    lemma="madame"/>
<w span="word_28"   msd="SBP"   lemma="vauquer" />
<w span="word_29"   msd=" "  lemma="," />
<w span="word_30"   msd="ADJ2PAR:f:s"   lemma="naître" />
<w span="word_31"   msd="PREP" lemma="de" />
<w span="word_32"   msd="SBP"   lemma="conflans" />
<w span="word_33"   msd=" "  lemma="," />
<w span="word_34"   msd="ECJ:3p:s:pst:ind"     lemma="être:3g" />
<w span="word_35"   msd="DTN:m:s"   lemma="un" />
...
```

Figure 5. *« Goriot » (extrait) : analyse morpho-syntaxique (WinBrill, sortie XML)*

Syntaxe	
S:prop("Madame_Vauquer")	Madame_Vauquer
,	
Vm:v-pcp2('naître' F S)	née
DN:pp	
=H:prp("de")	De
=DP:prop("Conflans")	Conflans
,	
Vm:v-fin('être' PR 3S IND)	est
Cs:np	
=DN:art('une' <idf> F S)	une
...	

Figure 6. *« Goriot » (extrait) : analyse syntaxique (VISL-FrAG, Bick 2003)*

Syntaxe et sémantique

```
prop. principale
        * Sujet [est]
                Madame  (polysémique : titre donné aux femmes nobles)
                Sous-groupe+
                        Vauquer
                        ,
                        née  (polysémique : venir au monde)
                        * Complément d'objet indirect [née]
                                de
                                Conflans
...
```

Figure 7. *« Goriot » (extrait) : analyse syntaxique et sémantique (Cordial)*

Coréférence

Madame Vauquer, née De Conflans , est une vieille femme qui, depuis quarante ans, tient à Paris <coref id="1">une pension bourgeoise</coref> établie rue Neuve-Sainte-Geneviève , entre le quartier latin et le faubourg Saint-Marceau . <coref id="2" type="ident" ref="1">Cette pension</coref>, connue sous le nom de la Maison-Vauquer , admet également des hommes et des femmes, des jeunes gens et des vieillards, sans que jamais la médisance ait attaqué les mœurs de ce respectable établissement.

Figure 8. *« Goriot » (extrait) : annotation coréférentielle (Vieira et al., 2005)*

2.2. Articulation avec les métadonnées

Les métadonnées associées à des données linguistiques en documentent le contenu, le format, l'historique de sa constitution et de ses révisions ainsi que les conditions de distribution. L'objectif principal de cette documentation est de faciliter l'identification et l'accès à la ressource, que ce soit à travers une interface de requête associée à une base de corpus ou par un référencement sur un serveur de métadonnées, tel que le serveur OLAC[8]. Contrairement aux données, qui peuvent éventuellement être soumises à des restrictions d'accès, les métadonnées sont, par principe, systématiquement libres d'accès. Parmi les initiatives majeures visant à inventorier les besoins en métadonnées spécifiques aux corpus linguistiques, la définition de l'en-tête de la TEI a joué un rôle précurseur. D'autres projets ont relayé l'initiative, en particulier OLAC, IMDI[9] et INTERA, avec la proposition d'un ensemble de métadonnées adaptées à divers types de ressources linguistiques (corpus oraux, écrits, multimodaux et lexiques) à partir du Dublin Core[10].

À l'organisation architecturale de la *FReeBank* en corpus, niveaux de description linguistique et ressources s'ajoute donc une documentation sous forme de méta-données. Ces métadonnées ont été choisies parmi celles proposées dans le cadre des initiatives mentionnées précédemment, en fonction de leur pertinence vis-à-vis des composantes de l'architecture sous-jacente et reposant crucialement sur la définition que nous avons donnée à chacune des composantes dans la section 2.1.

La composante *corpus* est caractérisée par les métadonnées permettant de documenter la couverture linguistique. Il s'agira d'informations globales relatives au contenu linguistique. Parmi celles-ci, mentionnons la (ou les) langue(s), des statistiques lexicales élémentaires (nombre de mots, de tours de paroles), ainsi que la caractérisation effective de la source, qu'il s'agisse de données écrites (ouvrage d'origine) ou orale (caractérisation des conditions de recueil).

La composante *niveau de description* est caractérisée par des métadonnées permettant de qualifier les choix linguistiques ou éditoriaux correspondant à la contribution spécifique du niveau de description, sans dupliquer ce qui a été décrit au niveau du corpus. Il pourra s'agir des choix de transcription, d'un schéma d'annotation particulier, de l'identification de la personne ou de l'outil à l'origine de ce niveau de description, et, le cas échéant, du niveau de description sur lequel s'ancre éventuellement le niveau courant (cas d'une annotation externe). En lien avec la gestion de différentes versions d'un même niveau de description, certaines de ces métadonnées sont commentés de façon plus détaillée dans la section 3.2.1.

Enfin, la documentation de la composante *ressource* sera limitée à l'identification de l'entité responsable du dépôt de la ressource en question, des dates de dépôt, ainsi que des conditions de diffusion spécifiques associées.

[8] http://www.language-archives.org/
[9] http://www.mpi.nl/IMDI/
[10] http://dublincore.org/

Symposium EPAL

2.3. Modélisation d'une archive de corpus linguistiques

Nous proposons maintenant d'intégrer les différents éléments présentés dans les sections précédentes au sein d'un modèle unique qui serve de base de représentation, mais aussi de déploiement, d'une archive ouverte de corpus linguistiques. Nous souhaitons ainsi, au-delà de l'expérience de la *FReeBank*, offrir un cadre conceptuel intégré de descriptions d'archives linguistiques au sens large. Le modèle, schématisé dans Figure 9, s'articule autour des trois composantes que sont le *corpus*, le *niveau de description* et la *ressource*. L'organisation de ces trois composantes entre elles fait apparaître de fait deux vues complémentaires attachées à la notion de corpus : une vue conceptuelle d'abord, qui permet de voir un corpus comme formé d'un certain nombre (*0* à *n*) niveaux de descriptions, identifiant l'organisation des différentes représentations linguistiques (description et/ou analyse du matériau linguistique) ; une vue opérationnelle ensuite, qui décompose le corpus en ressources (*0* à *n*), correspondant aux unités de dépôt et d'archivage.

Dans ce cadre, le corpus est le point d'entrée unique à l'intérieur de l'archive et peut, comme on le voit, n'être associé à aucune ressource ou niveau de description. On modélise ainsi le cas particulier d'un serveur de méta-données, qui ne fait que répertorier des ressources existantes, numérisées ou non (comme, par exemple, le portail OLAC). Du point de vue des niveaux de description, le modèle intègre la possible dépendance d'un niveau donné vis-à-vis d'un ou de plusieurs autres niveaux. Cette dépendance permet, comme nous l'avons vu, de gérer la logique d'organisation de certaines informations linguistiques entre elles, mais aussi de donner une base solide à la définition du caractère primaire ou secondaire d'un niveau de description donné.

Par ailleurs, le lien entre niveau de description et ressource explicite la contrainte qu'il ne puisse y avoir de ressource déposée qui ne contienne au moins un niveau de description. A l'inverse, nous n'imposons pas qu'un niveau de description soit explicitement associé à une ressource effective, considérant qu'une base de méta-données pourrait identifier des annotations disponibles pour un corpus donné, sans pour autant les archiver physiquement.

Enfin, la décomposition en niveaux de description d'un corpus donné est indépendante de l'intégration ou non de tout ou partie de ces niveaux de description en une seule et même représentation (paradigme de l'annotation interne). Quelle que soit la représentation adoptée, nous faisons l'hypothèse que les niveaux de description concernés sont clairement identifiés et documentés en tant que tel.

Combinées aux composantes du modèle d'architecture, les métadonnées définissent un modèle de documentation complet associé à la *FReeBank*, dans la ligné des travaux de modélisation menés au sein du TC 37/SC 4 (Ide et Romary, 2004a). D'un point de vue opérationnel, ces métadonnées se présentent sous forme d'en-têtes TEI, générés automatiquement à partir de l'interface de saisie lors du dépôt d'une ressource. Un certain nombre de métadonnées sont renseignées par le

Symposium Final

dépositaire, d'autres seront complétées par des informations générées automatiquement par la base (dates de dépôt d'une ressource, taille d'un corpus, statistiques sur les unités de segmentation élémentaires, etc.). Une version simplifiée de cette option est intégrée dans l'implémentation actuelle la *FReeBank*[11,12].

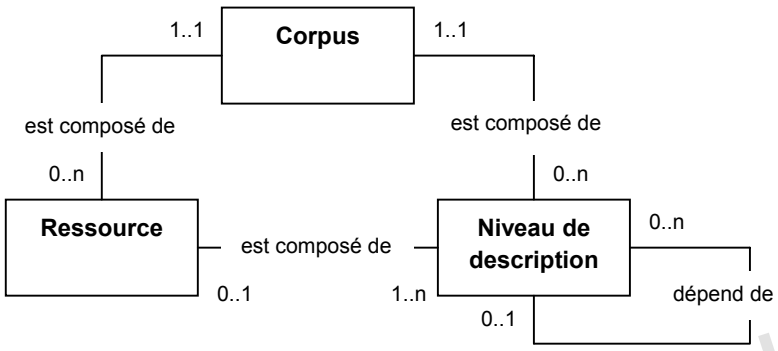

Figure 9. *Architecture pour une archive de corpus linguistiques*

Symposium Epal

3. Alimentation de la *FReeBank*

3.1. *Corpus et niveaux de description disponibles*

Le prototype de la *FReeBank* accueille d'ores et déjà certains corpus collectés et annotés dans le cadre du projet *Ananas* (Salmon-Alt, 2002). Lors de cette collecte, l'accent a été mis sur des corpus appartenant à des genres variés, et dans la mesure du possible, accessibles librement à des fins de recherche : textes littéraires, journalistiques, techniques, scientifiques et administratifs. Actuellement, nous disposons de données d'environ un million de mots. Les niveaux de description associés aux corpus sont présentés de façon plus détaillée dans la suite : segmentation en unités de référence, analyse structurelle en unités textuelles

[11] http://www.atilf.fr/freebank

[12] Note aux relecteurs : Il existe un prototype d'implémentation de la *FReeBank*, accessible sur http://www.loria.fr/projets/freebank. Ce prototype fonctionne pour les opérations élémentaires (dépôt et téléchargement de corpus), mais n'intègre ni totalité des concepts développés ici, ni la totalité des ressources disponibles et décrites dans la section 3.1. Nous sommes en train de redéployer le site à l'ATILF et y publierons les ressources avant l'été 2005. En attendant, l'ATILF met à disposition toutes les ressources décrites sur simple demande.

fondamentales (textes, titres, paragraphes, phrases, divisions), double voire triple annotation morpho-syntaxique, analyse syntaxique et annotation de certaines anaphores. Le Tableau 1 donne une vue synthétique des données actuellement disponibles dans la *FReeBank*. (cf. note 12).

Titre du corpus *Taille en mots* *Genre*	*Description*	*Constitution et contributions majeures (Institution, projet, personnes)*	*Niveaux de description disponibles*
Père Goriot 100.000 littéraire	extraits « Le Père Goriot » de H. Balzac	LORIA/LED (F. Bruneseaux)	TEI, segmentation, morpho-syntaxe, syntaxe
Vittoria 10.000 littéraire	« Vittoria Accoramboni, duchesse de Bracciano » (H. Beyle)	LIMSI (A. Popescu-Belis, I. Robba)	TEI, segmentation, morpho-syntaxe, syntaxe, anaphores
Alice 30.000 littéraire	extraits « Alice au pays des merveilles » (L Carroll)	LORIA/LED (Silfide)	TEI, segmentation, morpho-syntaxe
Les Misérables 100.000 littéraire	extraits « Les Misérables » (V. Hugo)	ATILF (Frantext)	TEI, segmentation, morpho-syntaxe
Le Monde 1997 65.000 journalistique	extraits « Le Monde » (09/1987)	ATILF (Parole) LORIA/LED (H. Manuélian)	TEI, segmentation, morpho-syntaxe, syntaxe, anaphores
Le Monde Diplo 120.000 journalistique	extraits « Le Monde Diplomatique » (1998)	LORIA/LED (Silfide) U de Grenoble 3 (C. Clouzot)	TEI, segmentation, morpho-syntaxe, anaphores
Est Républicain 20.000 journalistique	extraits « L'Est Républicain » (2001)	L'Est Républicain ATILF (Ananas)	TEI, segmentation, morpho-syntaxe
JOC 60.000 administratif	« Journal officiel de la CE » (questions-réponses des parlementaires, 1992)	LORIA/LED ATILF (CommonRefs)	TEI, segmentation, morpho-syntaxe, syntaxe, anaphores
Constitution 7.000 administratif	Constitution de la France (version du 4/10/1958)	ATILF (Ananas)	TEI, segmentation, morpho-syntaxe
Charte Dentistes 7.000 administratif	Charte professionnelle des chirurgiens-dentistes	ATILF (Ananas)	TEI, segmentation, morpho-syntaxe
MAIF 7.000 administratif	constats d'accidents de voiture	transmis par GREYC (P. Enjalbert)	TEI, segmentation, morpho-syntaxe, syntaxe, anaphores
Journal CNRS 25.000 scientifique	extraits du « Journal du CNRS » (CNRS Editions)	ATILF (Ananas)	TEI, segmentation, morpho-syntaxe

Table 1. *Corpus disponibles dans la FReeBank (avant l'été 2005, cf. note 12)*

Symposium Epal

3.1.1. Segmentation en unités de référence

L'objectif d'une première étape de traitement était de constituer les données correspondant au niveau de description primaire : il s'agit de données de référence, permettant de reconstituer la couverture des corpus et fournissant le point d'ancrage pour les autres niveaux de description. Les corpus actuels de la *FReeBank* fournissent à ce niveau un découpage en unités linguistiques minimales (Figure 3). Les composants des lexèmes complexes, des noms propres et des déterminants contractés ont été séparés. Ce choix aura permis d'ancrer de façon optimale les informations issues d'analyses ultérieures par une identification précise des unités linguistiques en question : une analyse morphologique qui décompose les déterminants (*au* en *à* + *le*) trouvera les deux points de référence nécessaires, alors qu'il n'y aura pas de perte d'informations dans le cas contraire. De même, au niveau syntaxique, ce choix permettra de distinguer avec précision le début d'un groupe prépositionnel de celui du groupe nominal imbriqué ([*à* [*le château*]]). Ces deux aspects se révèleront particulièrement importants pour l'annotation anaphorique. (section 3.1.5.)

3.1.2. Niveau de description structurelle

Nous ne souhaitions perdre aucune des informations présentes dans les corpus de départ, qu'il s'agisse d'informations linguistiques ou d'informations inhérentes à la structuration du texte, essentiellement en phrases, paragraphes, sections et titres (Figure 2). A partir d'un inventaire des différents éléments structuraux dans les corpus de la base, nous avons défini des schémas d'annotation suivant les recommandations de la TEI. Pour les ressources sans marquage explicite de la structure initiale, nous avons généré la description des principales unités structurelles repérables de façon automatique : sections, paragraphes, titres, phrases. La rétro-conversion des corpus qui étaient déjà annotés en paragraphes, tours de parole, titres, sections et/ou discours directs comporte en plus une phase de re-synchronisation des pointeurs avec les unités de référence (section 3.1.1.), travail qui est actuellement en cours.

3.1.3. Niveau de description morpho-syntaxique

Tous les corpus ont fait l'objet d'une annotation morpho-syntaxique. Ce niveau de description identifie des unités linguistiques pertinentes d'un point de vue morpho-syntaxiques (égaux ou supérieurs aux unités de référence décrits en 3.1.1.) et leur attribue une catégorie grammaticale, des informations flexionnelles et un lemme. Par défaut, ces informations sont celles issues d'une analyse effectuée par *Cordial*, dont le résultat a été converti en annotation XML externe (sur le même principe que celle de la Figure 5). Le lien avec les segments de référence se fait grâce à un pointeur sur un ou plusieurs de ces segments. Le cas d'une référence à plusieurs segments se présente lorsque plusieurs unités minimales ne forment

Symposium Epal

qu'une seule entité morphologique, par exemple pour les déterminants contractés ou des mots composés.

En plus de l'analyse effectuée par *Cordial*, nous avons sauvegardé des annotations morpho-syntaxiques éventuellement préexistantes. Par ailleurs, un étiquetage supplémentaire – basé sur le *DecisionTreeTagger* (Schmid 1994) et des grammaires locales de désambiguïsation (Bick 2003) – est réalisé au cours de l'analyse syntaxique. De ce fait, le niveau de description morpho-syntaxique de la *FReeBank* est l'illustration par excellence de l'utilité de la mise oeuvre du principe de représentation externe. Sans cette solution, la maintenance de plusieurs versions d'annotation morpho-syntaxique pour un même corpus aurait automatiquement nécessité la maintenance de plusieurs niveaux de description primaires, avec, comme conséquence majeure, la perte de la garantie de la cohérence de la couverture linguistique, et donc, in fine, de l'identité du corpus.

3.1.4. Niveau de description syntaxique

L'annotation syntaxique était une étape essentielle et critique, puisqu'elle était destinée à permettre l'extraction automatique des groupes nominaux et pronominaux, intervenant ultérieurement dans l'annotation anaphorique. L'identification des constituants syntaxiques ainsi que de leurs fonctions s'est faite à l'aide de l'analyseur *VISL-FrAG* (Bick, 2003), accessible librement en ligne[13]. Il s'agit d'un système multi-niveaux hybride, combinant approche probabiliste, grammaire à base de contraintes et grammaire de structures phrastiques. L'entrée, sous forme de texte brut, est d'abord étiquetée par le *DecisionTreeTagger,* puis corrigée et désambiguïsée à l'aide d'un lexique morphologique et de règles locales contextuelles. La sortie est traitée par un système hiérarchique de grammaires de contraintes, ajoutant et désambiguïsant des étiquettes pour les formes et fonctions syntaxiques. Ensuite intervient une grammaire de structures phrastiques, basée non pas sur des terminaux traditionnels (mots), mais sur les fonctions syntaxiques. L'avantage d'un tel système hybride est de combiner la robustesse des grammaires par contraintes avec la profondeur des grammaires de structures phrastiques. Afin de réduire les ambiguïtés liées aux structures coordonnées et au rattachement des groupes nominaux, une grammaire spécialisée pour les attachements est utilisée à un niveau intermédiaire. La dernière étape, la sélection automatique de l'arbre d'analyse, est optionnelle et peut être replacée par un choix semi-manuel. Pour l'instant, l'annotation syntaxique a été réalisée pour un quart des corpus disponibles de la *FReeBank*. Un dixième de ces données (20.000 mots environs) a d'ailleurs fait l'objet d'une correction et d'une validation manuelle (section 3.2.1).

3.1.5. Niveau de description coréférentielle et anaphorique

L'annotation de la coréférence et des anaphores demande d'abord un marquage des expressions entrant potentiellement dans des relations coréférentielles ou anaphoriques. Cela concerne les expressions (pro)nominales, mais aussi d'autres

[13] http://sandbox.visl.sdu.dk/visl/fr/

Symposium EPAL

types de constituants tels les groupes verbaux, les phrases ou même les paragraphes. En pratique, nous avons opté pour une extraction de tous les groupes (pro)nominaux de longueur maximale, puis un filtrage, excluant essentiellement les pronoms personnels autres que ceux de 3^{e} personne, possessifs, explétifs, vides, réflexifs, ainsi que les groupes nominaux appositifs, les noms suivant directement une préposition (*de cyclisme, en juillet*) et les expressions nominales temporelles (*ce matin, le lendemain*). Pour les corpus n'ayant pas fait l'objet d'une analyse syntaxique préalable, l'ensemble de ces constituants a été sélectionné manuellement. Pour les autres corpus, ces informations ont été extraites automatiquement, puis filtrées semi-automatiquement.

La deuxième étape a été l'annotation manuelle des liens coréférentiels et anaphoriques. Après plusieurs phases d'expérimentation sur différents corpus et avec différents annotateurs, nous avons obtenu le meilleur taux d'accord entre annotateurs en procédant par une annotation en plusieurs passes. D'abord, nous avons demandé aux annotateurs de séparer les anaphores coréférentes des autres anaphores. Ensuite, les anaphores coréférentes « fidèles » (à tête identique) ont été séparées des anaphores coréférentes « infidèles » (à tête divergente). Enfin, il s'agissait d'identifier, parmi les anaphores non coréférentes, celles ayant une relation bien typée avec l'antécédent (tout-partie, ensemble-membre, etc.).

Actuellement, les données de la *FReeBank* annotées en anaphores et/ou coréférence approchent les 180.000 mots (répartis sur neuf corpus). Elles comportent en tout 18.000 expressions référentielles et 8.500 liens référentiels. Environ la moitié de ces données provient de projets antérieurs (Bruneseaux 1997, Popescu-Belis et al. 1998, Clouzot et al. 2000) rétro-converties, l'autre moitié ayant été annotée par nos soins (Vieira et al., 2005).

3.2. *Questions de « bonne pratique » éditoriale*

La constitution des données actuelles de la *FReeBank* a été accompagnée par des décisions éditoriales concernant, d'une part, la validité linguistique des annotations, et d'autre part, la normalisation de leur représentation. Sans vouloir imposer nos choix à d'autres contributions venant potentiellement alimenter la *FReeBank*, nous considérons ces questions comme importantes lors de la mise en oeuvre d'une archive de ressources linguistiques réutilisables. C'est aussi dans cet esprit que nous avons conçu le prototype de la *FreeBank* comme une plate-forme de distribution de corpus à vocation générique et ouverte, et que le modèle formel d'architecture, développé dans la section 2., nous fournit le point de départ pour l'implémentation d'un portail évolutif, capable d'intégrer et de documenter des contributions de natures très diverses soumises en ligne, qu'il s'agisse de nouveaux corpus ou de mises à jour du matériau existant, d'annotations internes ou externes, ou de formats normalisées ou propriétaires.

Symposium Epal

3.2.1. Des contraintes sur la validité linguistique à la gestion des versions

En prolongement du débat sur le degré d'interprétation dans une annotation de corpus (cf. la discussion de la section 2.1.2.) se pose la question de la validité linguistique des données déposées à la *FReeBank*. Ayant conscience que la frontière entre variante interprétationnelle et analyse linguistique invalide peut être floue dans certains cas, nous proposons toutefois de faire reposer cette distinction sur l'accord entre jugements humains : si, pour une variante interprétationnelle, il peut y avoir divergence entre plusieurs jugements humains ou hésitation pour un même annotateur, une analyse linguistique fausse sera identifiée en tant que telle de façon convergente. Le premier cas est illustré par l'annotation anaphorique (Figure 10) : ici, un même annotateur humain hésite entre l'attribution de deux antécédents différents (*les technologies de l'information* vs. *les technologies de l'information* et *l'infosphère*) à une même anaphore pronominale (*elles*). La Figure 11 montre un cas de fausse analyse morphologique de la forme *suis*, analysée en tant que forme fléchie de *suivre* plutôt que de *être*.

```
Incertitude technique, d'abord, tant il est difficile de discerner les implications à moyen et long terme
de l'explosion <referentialMarkable id="m_1">des technologies de l'information </struct> et de
l'émergence d'<referentialMarkable id="m_2">une infosphère</struct> dont M. de Saint-Germain
souligne - incertitude supplémentaire - qu'<referentialMarkable id="m_3">elles</struct> sont pilotées
par le marché civil.
<alt>
    <referentialLink referentialSource="id(m_3)" referentialTarget="id(m_1),id(m_2)"/>
    <referentialLink referentialSource="id(m_3)" referentialTarget="id(m_1)"/>
</alt>
```

Figure 10. *Variante d'interprétation (« Le Monde Diplo », Clouzot et al., 2000)*

```
    <w lemma="ce">C'</w>
    <w lemma="être:3g">est</w>
    <w lemma="lui">moi</w>
    <w lemma="qui">qui</w>
    <w lemma="suivre:3g">suis</w>
    <w lemma="le">l'</w>
    <w lemma="auteur">auteur</w>
    <w lemma="de">de</w>
    <w lemma="ton">ta</w>
    <w lemma="joie">joie.</w>
```

Figure 11. *Analyse fausse (« Goriot », WinBrill)*

Le modèle d'architecture de la *FReeBank* étant générique, le degré de validité linguistique des données déposées dépend uniquement des décisions de nature éditoriales. Dans la perspective d'une plate-forme évolutive et ouverte à l'enrichissement progressif par des contributions de la communauté scientifique,

nous avons fait le choix de ne pas imposer une validation linguistique des données *a priori*. Cela signifie que certaines informations résultant d'une analyse linguistique humaine ou automatique sont potentiellement invalides ou sujet à des variantes interprétationnelles.

Concernant les informations invalides, il peut s'agir de fausses analyses automatiques, par exemple de certaines sorties d'un analyseur syntaxique, ou d'erreurs commises lors d'une annotation manuelle. Nous posons toutefois le postulat que les données déposées présenteront toujours de l'intérêt : la majorité des informations restera valide et exploitable, aucune annotation de corpus (ni automatique, ni humaine) n'est infaillible, et de la sorte, la *FReeBank* donnera une image relativement fidèle de l'état de l'art dans l'annotation automatique et manuelle de corpus. Par ailleurs, nous espérons qu'une dynamique d'évolution et d'amélioration, concernant à la fois les outils TAL et les ressources, s'installera à partir des données déposées et donc accessibles à l'ensemble de la communauté : repérage et correction automatiques des analyses linguistiques insuffisantes ou invalides, travail sur des mesures d'évaluation pour l'accord inter-annotateur, fusion de résultats de plusieurs analyses, études linguistiques manuelles sur certains phénomènes. Ce type de travaux pourra donner lieu à une stratification d'un même niveau de description : une analyse syntaxique automatique à large couverture pourra par exemple faire l'objet d'une correction manuelle selon le paradigme « transverse », c'est-à-dire focalisée sur certains points bien identifiés. Wallis (2003) a en effet constaté qu'une telle approche de la correction manuelle est à la fois moins coûteuse et plus fiable qu'une correction qui se veut exhaustive. Selon notre propre expérience, la correction manuelle exhaustive de la sortie de l'analyseur syntaxique *VISL-FrAG* (Bick, 2003) pour un extrait de 20.000 mots a demandé un homme/mois. Or, cette correction, bien qu'effectuée par une linguiste entraînée, a introduit de nouvelles erreurs. Cela signifie qu'il faudrait prévoir, pour aboutir à des données supposées « valides », au moins une double correction, une phase de comparaison systématique, une phase de concertation, puis une phase de prise de décision par une tierce personne en cas de désaccord subsistant. L'extrapolation de notre expérience à un corpus de 1 million de mots aboutirait alors à un coût approximatif de 150 hommes/mois, coût qui est par ailleurs réaliste comparé à celui du projet SALSA (annotation de rôles sémantiques), procédant précisément selon ces principes (Erk et Pado, 2004). En considérant que l'état de l'art est en évolution très rapide, il nous paraîtrait vain de mettre en œuvre de tels moyens dans l'espoir de créer un possible corpus de référence qui serait, à peine produit, immédiatement dépassé.

Une partie des données linguistiques de la *FReeBank*, même linguistiquement valide, sera toujours sujette à des variantes interprétationnelles. Bien que Leech (1993) préconise l'usage de schémas d'annotation aussi « neutres » que possibles par rapport à des théories particulières, on ne peut exclure des variantes dues à des approches théoriques divergentes. D'autres variantes proviennent plus simplement de différentes analyses automatiques (cf. le niveau d'annotation morpho-syntaxique

Symposium EPAL

des corpus noyau de la *FReeBank*) ou de la possibilité de plusieurs interprétations humaines concurrentes (« ambiguïtés »). Il s'agit là d'un phénomène tout à fait régulier pour les langues naturelles, qu'il n'y aurait aucun intérêt à exclure de la *FReeBank*. Au contraire, leur marquage explicite peut s'avérer extrêmement utile, par exemple pour la définition de clés lors de campagnes d'évaluation d'outils TAL ou pour la validation d'approches cognitives.

Qu'il s'agisse de données visant à « corriger » des données existantes ou de données sujettes à des interprétations multiples, la gestion de ce matériau dans la *FReeBank* passe obligatoirement par une documentation fine sous forme de méta-données appropriées (cf. section 2.2). Plus précisément, il s'agit de spécifier des liens entre données parallèles relevant à la fois d'un même corpus et d'un même niveau de description. Les principaux cas de figure relevés précédemment – correction exhaustive, correction transverse, co-existence de plusieurs analyses humaines ou automatiques – peuvent en effet se caractériser à partir d'une combinaison de métadonnées subordonnées au niveau de description. Parmi ceux-ci, deux revêtent d'une importance particulière : la granularité du niveau de description et la validation linguistique explicite.

La granularité d'un niveau de description caractérise la nature et l'étendue des phénomènes linguistiques pris en compte et s'exprime par exemple par référence à une DTD. Toutefois, elle devra être formalisée à terme sous forme d'une référence à une sélection de catégories de données du registre des catégories de données[14] (Ide et Romary, 2004a). Par rapport à des données préexistantes dans la *FReeBank*, une nouvelle soumission relative à ces données (même corpus, même niveau de description) peut se caractériser par exemple par une granularité égale, plus fine ou différente. À granularité égale, il s'agira du dépôt d'une *version parallèle* : c'est par exemple le cas d'un nouvel étiquetage morpho-syntaxique automatique ou d'une deuxième annotation manuelle des pronoms personnels anaphoriques. À granularité plus fine, les données soumises fourniront une *version parallèle enrichie* : par exemple une annotation morpho-syntaxique sous-catégorisant plus finement les adverbes ou une annotation anaphorique étendue aux descriptions définies. À granularité différente, il s'agit tout simplement de *versions supplémentaires* (non parallèles).

La validation, elle, consiste en l'affirmation explicite, de la part d'un annotateur, d'un dépositaire ou d'un administrateur de la *FReeBank*, de la validité linguistique de données soumises. Lorsque la couverture et le niveau de description des données soumises correspondent à une version préexistante de la *FReeBank*, une *version validée* peut constituer de fait une correction exhaustive ou transverse. Il s'agira d'une correction exhaustive, lorsque la granularité de la version validée est égale ou supérieure à la version précédente. A granularité inférieure ou différente, il s'agira d'une correction transverse.

[14] http://syntax.loria.fr

Copyright Eyrolles

Les métadonnées liées à la granularité des niveaux de description, combinées à la notion de validation linguistique explicite suffisent donc à caractériser tous les cas de co-existence de plusieurs analyses humaines ou automatiques ainsi que les cas des corrections exhaustives ou partielles relatives à des données initialement non validées. Nous pensons qu'offrir la possibilité de gérer convenablement ces variations est un élément essentiel pour qu'une archive telle que la *FReeBank* reflète au mieux les aspirations d'une large communauté scientifique.

3.2.2. Représentation normalisée des données linguistiques

Pour la représentation des données noyau de la *FReeBank*, nous nous sommes fixés comme priorité de garantir la compatibilité avec les standards internationaux, en mettant en oeuvre les recommandations pour la représentation de ressources linguistiques actuellement en cours de développement au sein du TC 37/SC 4 de l'ISO et relayées en France par l'initiative RNIL[15]. Celles-ci concernent plus particulièrement les niveaux de description correspondant aux « annotations » linguistiques au sens de McEnery et Wilson (1996), introduites dans la section 2.1. Dans cette optique, le processus d'annotation linguistique consiste, schématiquement, en l'identification d'unités pertinentes d'un point de vue linguistique, puis éventuellement, en la caractérisation de ces unités par des traits linguistiques et/ou en l'établissement de liens entre ces unités. Le résultat d'une annotation linguistique se présente dans un format particulier (parenthésage, base de données relationnelle, balisage etc.). Or, certaines initiatives précédentes, par exemple issues des campagnes d'évaluation américaines (Gerber et al., 2002), ayant tenté de proposer des standards sous forme de jeux de balises, se sont heurtées à un inconvénient majeur qui fut le manque de flexibilité, puisque l'on imposait aux utilisateurs à la fois le modèle de données sous-jacent (définition des propriétés des unités et des liens) et le format de représentation (SGML/XML+DTD).

Dans la lignée d'autres initiatives, plus génériques (Mengel et all., 2000; Bird et Liberman, 2001), le *Linguistic Annotation Framework* (LAF), tel que défini par Ide et Romary (2004b) préconise une séparation claire entre le modèle des données et les formats de représentation, partant du principe que la standardisation de l'annotation linguistique doit s'effectuer au niveau conceptuel plutôt qu'au niveau représentationnel. LAF propose donc une modélisation conceptuelle des objets d'annotation sous forme d'un méta-modèle. Celui-ci reflète les propriétés structurelles des données d'annotation : il s'agit d'un graphe orienté représentant les d'unités pertinentes d'un point de vue linguistique ainsi que les contraintes régissant leur agencement. Les propriétés de ces unités sont caractérisées par des descripteurs linguistiques ou « catégories de données », dont la gestion est externalisée dans un registre de catégories de données en ligne (Ide et Romary, 2004a; cf. note 14). L'association d'un ensemble de catégories de données aux nœuds d'un méta-modèle donne lieu à un modèle de données pleinement spécifié. Si LAF ne se préoccupe pas

[15] http://pauillac.inria.fr/atoll/RNIL/home-fr.html

Symposium EPAL

prioritairement des formats d'instanciation (i.e. des schémas d'annotation concrets), il propose néanmoins un format de représentation pivot parfaitement isomorphe au modèle. Sera considérée comme étant conforme à LAF toute annotation pour laquelle il existe une procédure d'appariement avec ce format générique.

L'élaboration de modèles de données est actuellement en cours pour différents niveaux de description[16] : morpho-syntaxe (Clément et de la Clergerie, 2004) syntaxe (Brants et al., 2002, Ide et Romary, 2003), rôles sémantiques (Erk et Pado, 2004), coréférence et anaphores (Salmon-Alt et Romary, 2005). Pour chaque niveau, il s'agit d'identifier le méta-modèle et l'ensemble de catégories de données spécifiques à la description linguistique de ce niveau. Pour les données de la *FReeBank*, nous avons suivi ces initiatives, en essayant de les mettre en oeuvre de façon aussi systématique que possible.

Le TC 37/SC 4 travaille dès à présent à la définition d'un modèle générique dédié à l'annotation morpho-syntaxique (future norme ISO 24611 ; Clément et de la Clergerie, 2004). Ce modèle combine d'une part deux niveaux de segmentation et de catégorisation linguistique, et d'autre part un ensemble de catégories de données linguistiques permettant de qualifier les différents éléments du modèle. Une étude préliminaire a en particulier permis de regrouper une base de catégories morpho-syntaxiques intégrant la plupart des jeux d'étiquettes connus pour le français. L'expérience d'appariement de ces principes avec les données de la *FreeBank* a été particulièrement intéressante, puisque elle a permis de mettre à l'épreuve les recommandations sur les sorties de quatre formats morpho-syntaxiques différents. Les propriétés structurelles ainsi que les descripteurs du format retenu pour l'encodage homogène des informations morpho-syntaxiques de la *FReeBank* sont compatibles avec ces principes et permettront aux données d'être converties vers la future norme sans perte d'informations.

Concernant le niveau de description syntaxique, la sortie de l'analyseur était initialement conditionnée par les fondements théoriques du projet *VISL*[17]. Or, s'agissant d'un environnement multilingue pour l'analyse et l'apprentissage syntaxiques, les structures et descripteurs syntaxiques sous-jacents (« méta-modèle » et « catégories de données ») se sont révélés suffisamment génériques pour avoir été mis à l'épreuve sur d'autres langues (portugais, danois). Par ailleurs, le format de sortie de l'analyse a pu être converti de façon complètement automatique vers le format *TIGER*. Il s'agit là de l'un des formats convergents au niveau international pour l'encodage des formes et fonctions syntaxiques (Brants et al., 2002) que nous avons retenu pour la *FReeBank*, pour plusieurs raisons : il est conçu pour représenter les résultats d'une analyse syntaxique profonde (dépendances et constituants), il permet la mise en oeuvre d'une annotation externe

[16] L'annotation structurelle étant couverte par la TEI (Sperberg-McQueen et Burnard, 2002).
[17] http://beta.visl.sdu.dk

Symposium EPAL

et il est accompagné par des outils de recherche de corpus arborés performants et libres[18].

Enfin, la prolifération des schémas pour l'annotation de la coréférence et des anaphores a donné lieu, depuis quelques années, à des tentatives de synthèse (Poesio, 2000 ; Salmon-Alt, 2001). L'idée-clé de ces initiatives est de proposer un modèle de données qui fédère les propriétés des schémas précédents et qui puisse être instancié selon les besoins concrets du codeur. Les grandes lignes de ces propositions – introduction d'éléments autonomes pour les expressions à annoter ainsi que dles liens entre celles-ci – ont été intégrées dans un travail visant la normalisation du codage de ce niveau de description (Salmon-Alt et Romary, 2005). La *FReeBank* sert actuellement de banc d'essai pour l'application des ces propositions à large échelle : toutes les annotations anaphoriques et coréférentielles existantes ont été converties automatiquement vers le modèle en cours de normalisation[19] et les catégories de données pertinentes pour la description des expressions ainsi que des liens ont été soumises au registre des catégories de données.

4. La *FReeBank* : une vitrine des bonnes pratique et un espace de réflexion ?

L'objectif clairement affiché des travaux que nous avons menés au cours des derniers mois autour du projet *FReeBank*, tant d'un point de vue théorique que de la réalisation effective d'une première plate-forme, est d'arriver à stabiliser un certain nombre de concepts fondamentaux reflétant l'état de l'art dans les domaines de la représentation et de l'archivage de corpus linguistiques. Cette vision intégrée de la création, de la gestion et de la diffusion de telles archives doit maintenant être relayée au sein de notre communauté, par des actions plus spécifiques dans différentes directions :

- contribution à l'édification et à la diffusion des normes et bonnes pratiques de représentation de données linguistiques : qu'il s'agisse de la TEI (dont la France est maintenant l'un des quatre sites hôte) ou de l'ISO (avec le rôle moteur du réseau RNIL), nous devons déterminer des actions (comparaison des pratiques existantes, tutoriaux, etc.) qui nous permettent de faire vivre les normes existantes et de les faire évoluer ;
- production d'échantillons reflétant ces bonnes pratiques : un travail concerté de différentes équipes de recherche devrait fournir des jeux de données annotés qui, dans la continuité des projets Ananas ou Asila, puissent servir d'exemple pour de nouveaux projets similaires ;
- identification d'un vrai champ de recherche autour des ressources linguistiques : il apparaît que la réponse aux problèmes de la gestion d'annotations multi-niveau ne peut se contenter d'une approche strictement

[18] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
[19] Les outils de conversion sont en libre accès sur http://www.atilf.fr/ananas.

Symposium - Epal

technologique. Les futurs travaux en la matière doivent s'appuyer sur des approches conceptuelles solides, qui dépassent la simple connaissance d'XML ;

– édification de principes généraux concernant l'archivage ouvert de ressources linguistiques : en espérant que d'autres projets d'archives linguistiques ouvertes voient le jour, il faut définir dès à présent des principes techniques minimaux en garantissant l'interopérabilité (interrogation multi-archive), mais surtout édifier une charte de telles archives qui puisse synthétiser l'opinion de notre communauté en matière de libre accès.

Les propositions faites ici ne sont qu'une étape pour avancer dans ces différentes directions et de fait, nous entrevoyons dès à présent de refondre complètement notre prototype initial (cf. note 12) dans une plate-forme qui, d'une part, intègre mieux le point de vue des utilisateurs dans l'organisation des mécanismes de dépôt et d'accès, et, d'autre part, s'articule plus étroitement avec les modèles de données associés aux différents niveaux de description (tant pour l'écrit que pour l'oral).

Bibliographie

Anderson A., Bader M., Bard E., Boyle E., Doherty G., Garrod S., Isard A., Kowtko J., McAllister J., Miller J., Sotillo S., Thompson H., Weinert R., "The HCRC MapTask Corpus", *Language and Speech*, vol. 34 no. 4, 1991, p. 351-366.

Bick E., "A CG & PSG Hybrid Approach to Automatic Corpus Annotation". *Proceedings of SproLaC*, Lancaster, Royaume-Uni, 2003.

Bird S., Liberman M., "A formal framework for linguistic annotation", *Speech Communication*, vol. 33 n° 1,2, 2001, p. 23-60.

Brants T., Hansen S., "Developments in the TIGER Annotation Scheme and their Realization in the Corpus", *Proceedings of LREC 2002*, Las Palmas, Espagne, 2002.

Bruneseaux F., Romary L., "Codage des références et coréférences dans les DHM", *Actes de ACH-ALLC '97*, Kingston Ont., Etats-Unis, 1997.

Clément L., de la Clergerie E., "Language Resource Management : Morpho-Syntactic Annotation", Working Draft ISO TC37/SC4/WG2, Jeju, Corée, 2004.

Clouzot C., Antoniadis G., Tutin A., "Constitution and Exploitation of an Annotation System of Electronic Corpora: Toward Automatic Generation of Understandable Pronouns in French Language", in *Lectures Notes in Artificial Intelligence 1835,* Christodoulakis D. (ed.),2000, p. 242-252.

Erk K., Pado S., "A powerful and versatile XML Format for representing role-semantic annotation", *Proceedings of LREC 2004*, Lisbonne, Portugal, 2004.

Gerber L., Ferro L, Mani I, Sundheim B., Wilson G., Kozierok R., "Annotating Temporal Information: From Theory to Practice.", *Proceedings of the 2002 Conference on Human Language Technology*, San Diego, CA, 2002, p. 226-230.

Habert B., Nazarenko A., Salem A., *Les linguistiques de corpus*. U Linguistique, Paris, Armand Colin/Masson, 1997.

Ide N., Priest-Dorman, G., *The Corpus Encoding Standard*, http://www.cs.vassar.edu/CES, 1996.

Symposium Epal

Ide N., Romary L., "A Registry of Standard Data Categories for Linguistic Annotation", *Proceedings of LREC 2004*, Lisbonne, Portugal, 2004(a).

Ide N., Romary L., "International standard for a linguistic annotation framework", *International Journal of Natural Language Engineering*, vol. 10 n° 3-4, 2004 (b), p. 211-225.

Ide N., Romary L., "Encoding Syntactic Annotation", in *Treebanks, Building and Using Parsed Corpora.* A. Abeillé (ed.), Kluwer Academic Publishers, Dordrecht, Boston, London, 2003.

Ide N., Véronis J. (eds.), "The Text Encoding Initiative: background and context", *Special issue of Computers and the Humanities*, vol. 29 n° 1/2/3, 1995.

Kilgarriff A., Grefenstette G., "Web as Corpus", Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics*, vol. 29 no. 3, 2003, p. 333-348.

Leech G., "Introduction to corpus annotation", in *Corpus annotation: linguistic information from computer text corpora*, Garside R., Leech G., McEnery T. (eds), Longman, London, 1997, p. 1-18.

Leech, G., "Corpus annotation schemes", *Literary and Linguistic Computing*, vol. 8 no. 4, 1993, p. 275-81.

McEnery T., Wilson A., *Corpus Linguistics*, Edinburgh, Edinburgh University Press, 1996.

Mengel, A., Dybkjaer, L., Garrido, J.M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C., *MATE Dialogue Annotation Guidelines*. 2000, http://www.ims.unistuttgart.de/projekte/mate/mdag/.

Poesio M., "Coreference", *MATE Dialogue Annotation Guidelines-Deliverable 2.1*., 2000, http://www.ims.unistuttgart.de/projekte/mate/.

Salmon-Alt S., "Du corpus à la théorie : l'annotation (co-)référentielle". *Traitement Automatique des Langues*, vol. 42 n° 2, 2001, Hermès, Paris, p. 459-487.

Salmon-Alt S., "Le projet ANANAS : Annotation Anaphorique pour l'Analyse Sémantique de Corpus", *Actes du Workshop sur les Chaînes de référence et résolveurs d'anaphores, TALN 2002,* 28/06/02, Nancy, 2002.

Salmon-Alt S., Romary L., "The reference annotation framework : a case of semantic content representation", *Proceedings of the International Workshop on Computational Semantics*, Tilburg, Pays-Bas, 2005.

Schmid H., "Probabilistic Part-of-Speech Tagging Using Decision Trees", *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 1994.

Sperberg-McQueen C.M., Burnard L. (eds.), *Guidelines for Text Encoding and Interchange (P4),* TEI Consortium, Humanities Computing Unit, University of Oxford, 2002.

Thompson H., McKelvie D., "Hyperlink semantics for stand-off markup of read-only documents", *Proceedings of SGML Europe '97*. 1997, Barcelona.

Véronis J., "Annotation automatique de corpus : panorama et état de la technique", in *Ingénierie des langues*, Pierrel J.-M. (ed), Paris, Hermes, 2000, p. 111-129.

Vieira R., Salmon-Alt S., Gasperin Caroline, Schang E., Othero G., "Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus", in *Anaphora Processing*. A. Branco, T. McEnery and R. Mitkov (eds.), John Benjamins Publishing Company, 2005.

Symposium Epal

Vieira, R., Gasperin, C. V., Goulart, R. V., “From manual to automatic annotation of coreference”, *Proceedings of the International Symposium on Reference Resolution and its Application on Question Answering Systems*, Venise, Italie, 2003, p.17-24.

Wallis S., “Completing Parsed Corpora, from Correction to Evolution”, in *Treebanks, Building and Using Parsed Corpora.* A. Abeillé (ed.), Kluwer Academic Publishers, Dordrecht, Boston, London.

Symposium Epal