Un besoin de spécifications des corpus de formation en ligne

Muriel Noras

Un besoin de spécifications des corpus de formation en ligne

Muriel Noras*

** LIFC – Université de Franche-Comté*
16 route de Gray
F-25030 Besançon cedex
muriel.noras@lifc.univ-fcomte.fr

RÉSUMÉ. *L'objectif de cet article est l'identification de méthodes de structuration des données pouvant s'appliquer aux corpus. Nous proposons donc un état de l'art des outils, des méthodes de transcription et/ou d'analyse des données ainsi que quelques spécifications de divers corpus. Par corpus, nous entendons un ensemble volumineux de données numériques hétérogènes (i.e., divers formats de fichiers) contenant, entre autres, les interactions et la situation d'apprentissage. Ceci dans l'optique de définir un format de structuration de corpus de formation en ligne qui soit commun à toutes les disciplines concernées par le domaine des EIAH (informatique, psychologie, sciences de l'éducation, pédagogie, sciences de l'information et de la communication, sociologie, etc). Ainsi, nous souhaitons favoriser la mise à disposition et les échanges de corpus.*

MOTS-CLÉS : *corpus, transcription, spécification, pluridisciplinarité, formation en ligne.*

1. Introduction

Dans le domaine des EIAH, les études concernant les travaux en groupe, i.e. travaux menés ollectivement/collaborativement/coopérativement, s'appuient sur des ensembles de données numériques issues d'expérimentations. Ces données, que nous appelons corpus, contiennent, entre autres, les interactions et la situation d'apprentissage. La diversité des formats de fichiers de données (fichiers audio, vidéos, images, textes, logs de clavardage, productions…) rend difficile l'exploitation des corpus par des outils informatiques. De plus, nous constatons que l'échange et la mise à disposition de corpus dans ce domaine restent encore peu fréquents pour diverses raisons telles que le refus de partager des ressources dont la collecte est le fruit d'un dur labeur, ou bien la difficulté de décrire le contenu des données à des chercheurs n'ayant pas participé à la formation. Par conséquent, tel que le montre [HENRI & CHARLIER 05], les recherches sur ces données empêchent la comparaison des outils, méthodes ou résultats d'analyse : les résultats publiés dans le champ des EIAH souffrent d'un manque de validité externe.

Il nous semble important de définir, dans un premier temps, un format commun de structuration des corpus afin de favoriser les échanges.

Les corpus auxquels nous nous intéressons proviennent de dispositifs de formation collaborative en ligne. Compte tenu de la difficulté d'exploitation de la multitude de fichiers aux formats hétérogènes, il est indispensable de leur donner une cohérence dans un corpus afin de faciliter l'automatisation des analyses. Nous souhaitons donc définir un format et une structuration de corpus suffisamment généraux pour permettre des analyses différentes par des chercheurs de disciplines diverses. Pour cela, il nous apparaît important d'étudier les méthodologies de transcription de données dans les corpus en se basant par exemple sur l'expérience du domaine du TAL (Traitement Automatique des Langues). Cette étude a pour but d'identifier les données transcrites et analysées ainsi que leurs formats et méthodes afin d'en tenir compte lors de nos définitions de format et de structuration. Le présent article propose un état de l'art de ces outils, des méthodes d'analyse et des spécifications des métadonnées ainsi qu'une définition de nos objectifs de recherche.

2. Environnements et méthodes de transcription et/ou d'analyse de données multimodales

2.1. Définition

Parmi les environnements de transcription/analyse existant, nous distinguons ceux concernant l'annotation de données, de ceux s'intéressant à leur transcription.

L'annotation revient à associer un contenu textuel secondaire à une portion de données (textuelles, vidéo, audio) étudiée. On considère que la valeur de cette annotation fait partie des données du corpus. Par exemple, si nous considérons l'énoncé de forum « La question 3 du sujet est-elle en relation avec la précédente ? », une annotation pourrait être « la question 3 : xxxx ?; la question 2 : yyyyy ? ». Cette forme d'annotation permet la désambiguïsation de l'énoncé.

La transcription est le processus permettant de restituer le plus objectivement possible, dans un format exploitable (format de fichier indexé par base de données ou autre) par les outils informatiques, un acte fondu dans un format brut (fichier vidéo, fichier audio, fichier texte). Nous appelons acte, une action réalisée par un individu : prise de parole, énonciation, écriture d'un message, vote, etc.

Selon la définition adoptée de l'annotation, celle-ci peut être soit objective soit subjective. Le processus de transcription de corpus doit être objectif dans la mesure où il ne doit pas être soumis à une quelconque interprétation. En revanche l'analyse, étape suivante du traitement des données recueillies, est, par nature, subjective. L'articulation de ces différentes phases est illustrée figure 1.

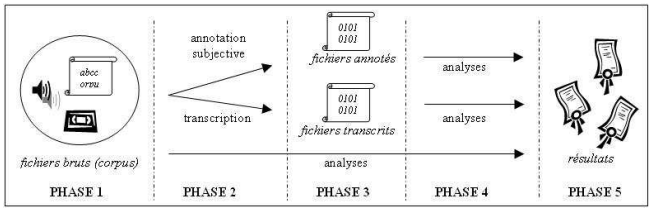

Figure 1. *Du fichier brut aux résultats de l'analyse*

Les environnements de transcription et/ou d'annotation de corpus vidéos, audios ou textuels conçus pour l'analyse de la collaboration émanent des travaux d'équipes des domaines de SHS (Sciences Humaines et Sociales) et EIAH. Par exemple : Praat, Transcriber et ELAN en SHS ; CoLAT en EIAH. Ils utilisent des corpus propres à leur domaine. Malgré l'existence de ColAT et ELAN, outils permettant l'analyse de plusieurs médias, peu d'outils sont réellement adaptés à l'analyse de corpus multimodaux.

2.2. Environnements

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* à Lyon [ISGS 06]. Les outils comparés étaient au nombre de six, à savoir Anvil, ELAN, EXMARaLDA, Media & Text Editors, TASX et MacVisSTA. Afin d'établir un tableau comparatif pertinent, les participants devaient effectuer les tests avec quatre corpus donnés (des vidéos) dont chacun des contenus différait (conversation libre, narration d'histoire, description

d'un itinéraire routier, travail collaboratif d'organisation). L'objet de ces tests était double : d'une part d'examiner comment gestuelle et discours varient selon les situations ; d'autre part d'établir un comparatif entre ces outils sur les mêmes données.

ANVIL est un outil de transcription développé par l'université de Sarre (Allemagne) dont l'objectif est de faciliter l'annotation et l'analyse des données verbales et non verbales dans le domaine des sciences du langage. Ces annotations figurent dans des lignes parallèles présentées sous forme de portées musicales. Une interface très ergonomique permet l'alignement temporel de divers types d'annotations et de graphiques avec le signal sonore et la vidéo. Le nombre de niveaux d'analyse, leurs relations hiérarchiques ainsi que les catégories de notation peuvent être configurés par l'utilisateur.

EXMARaLDA (EXtensible MARkup Language for Discourse Analysis), outil de transcription développé par l'université de Hambourg (Allemagne), a été conçu afin de transcrire les interactions verbales entre de nombreux interlocuteurs. Les interactions de comportements non-verbaux sont considérées comme des annotations des énoncés verbaux. L'outil reprend également l'idée de « partition » pour la présentation des transcriptions des paroles énoncées. Il semble pertinent si la transcription s'applique principalement aux actions verbales (paroles) ; en revanche son application pour la transcription en quantité égale d'actions verbales et non-verbales ne semble pas appropriée dans la mesure où l'interface est moins intuitive.

L'institut de psycholinguistique Max Planck aux Pays-Bas a créé l'outil ELAN (EUDICO Linguistic Annotator) afin de répondre à des besoins d'annotation et d'exploitation d'enregistrements multimédia. ELAN est spécialement conçu pour l'analyse du langage oral, du langage des signes, de la gestuelle, mais il peut également être utilisé pour toute annotation, analyse et documentation de corpus médiatiques (i.e., avec des données vidéo et/ou des données audio). ELAN est un outil d'annotation permettant de créer, éditer, visualiser et rechercher des annotations portant sur des données audio et vidéo. Contrairement à ANVIL et EXMARaLDA, il ne permet pas de mettre en relation les niveaux d'annotation.

L'outil ColAT (Collaboration Analysis Toolkit) est le fruit de recherches de l'équipe HCI (Human Computer Interaction) de l'université de Patras en Grèce. L'objectif visé est la réalisation d'un outil permettant l'analyse d'activités collaboratives utilisant des outils de communication médiatisée et engageant élèves et enseignants. De plus, cet outil doit permettre d'analyser l'ensemble des informations collectées au cours de l'activité de résolution de problème, autrement dit, une analyse multi-médias (e.g. fichiers de logs, vidéo/audio, texte, images). L'environnement permet d'effectuer une analyse de l'activité des acteurs selon différents points de vue, en utilisant le cadre de la théorie de l'activité. L'outil est destiné à des chercheurs et analystes en ethnographie afin de les aider à analyser des données collectées à partir de différentes sources. D'après [AVOURIS et al. 03],

quelques outils d'analyse existent mais ne permettent pas une analyse de plusieurs types de médias, simultanément ou non.

2.3. Méthodes

La méthodologie OCAF (Object-oriented Collaboration Analysis Framework) décrite dans [AVOURIS et al. 02] est une méthode dont le cadre conceptuel est la théorie de l'activité. Il s'agit d'une méthode proposée pour analyser et évaluer des situations collaboratives de résolution de problème, médiatisées par des supports aux activités collaboratives. Ce cadre d'analyse s'intéresse aux objets abstraits et concrets manipulés lors de l'élaboration d'une solution, alors que les techniques existantes sont davantage axées sur les stratégies et dialogues des acteurs. L'utilisation de la théorie de l'activité comme cadre conceptuel pour l'analyse des activités met en avant la possibilité de décomposer l'activité en plusieurs niveaux hiérarchiques (opérations, tâches, activité) et ainsi d'affiner le processus d'analyse de l'activité par plusieurs niveaux d'abstraction. Cette méthode se décompose en trois phases : définition du schème d'analyse des événements ; annotation des événements observés, contrôle et interprétation des données de l'activité ; analyse multi-niveaux de l'activité en établissant la correspondance opérations-actions.

MATE (Multi-level Annotation Tools Engineering) [HEID & MENGEL 99] est un projet dont l'objectif méthodologique est de faciliter la réutilisation des ressources d'un langage, notamment par la proposition d'un formalisme pour l'annotation à des niveaux différents et d'une boîte à outils supportant ce formalisme. Les niveaux de description se rapportent à l'analyse du langage, autrement dit, à la prosodie, la morphosyntaxe, les actes de dialogue etc. L'approche du projet MATE repose sur la définition d'un extrait de schème d'annotation pour chaque niveau linguistique permettant d'établir des liens entre plusieurs entités d'un même ou de différents niveaux. Le formalisme XML a été choisi pour des raisons de neutralité face au contenu et sa capacité à décrire n'importe quelle information.

Les environnements et méthodes de transcription décrits dans cette section nous permettent d'identifier les données transcrites et analysées, éléments importants pour notre définition de spécification. Il est également important de s'intéresser aux spécifications de métadonnées existantes pour, éventuellement, les réutiliser.

3. Spécifications des métadonnées

3.1. Aperçu général

Les spécifications ont pour objectif la définition de méthodes et modèles afin de faciliter l'échange de données, l'interopérabilité entre composants logiciels et la comparaison de résultats. En effet, l'utilisation d'un « langage/cadre » commun rend la communication plus aisée. La définition de spécifications des métadonnées

impose d'effectuer, préalablement, des choix concernant les informations à décrire dans les corpus, leur granularité et leur structuration.

3.2. Spécifications existantes (pour les corpus)

Dans le domaine des sciences humaines, la TEI (Text Encoding Initiative) [TEI 06] est un format d'échange de gros corpus de textes électroniques. Créée en 1987 par et pour des scientifiques intéressés par l'étude des textes (par exemple : spécialiste de littérature, historien, sociologue, linguiste, ethnologue, philosophe), son objectif est de permettre aux chercheurs d'échanger des corpus de texte mais aussi de mettre à leur disposition un système de codage et de balisage commun et normalisé. La structuration des textes a lieu par un balisage XML. De plus, la TEI permet d'intégrer des annotations et donc de les échanger.

L'ISO/TC37/SC4 [IDE & ROMARY 02] [TC37/SC4 06] est un groupe travaillant actuellement sur une spécification concernant la terminologie et d'autres ressources langagières de contenu. Plus précisément, le sous-comité SC4 s'intéresse à la gestion des ressources linguistiques. L'objectif est de préparer plusieurs standards/normes en spécifiant les principes et méthodes de création, codage, traitement et gestion des ressources langagières, tels que les corpus écrits, les corpus oraux, l'établissement de dictionnaires et les schèmes de classification.

La spécification LOM (Learning Object Matadata), centrée sur la FOAD (Formation Ouverte et A Distance), part des métadonnées du Dublin Core, basé sur documentation, en les étendant. Autrement dit, il permet de spécifier un ensemble de métadonnées propres au monde éducatif et de décrire les objets constituant les systèmes d'enseignement. Plus précisément, LOM permet de décrire les caractéristiques pédagogiques des ressources et ainsi de les indexer [DE LA PASSARDIERE & JARRAUD 05].

L'IMS (Instructional Management Systems), créé en 1997, rassemble des universités et des entreprises américaines. Il reprend l'essentiel de la LOM.

IMS Learning Design [KOPER & TATTERSALL 05] est une spécification d'IMS basée sur le langage de modélisation pédagogique EML (Educational Modelling Language). L'objectif du projet IMS-LD est de développer un environnement de modélisation d'unités d'apprentissage supportant la diversité, l'innovation et les différentes approches pédagogiques, promouvant l'échange et l'interopérabilité des « matériels pédagogiques », permettant de décrire des situations mono ou multi utilisateurs, individuelles ou collectives, en présentiel ou à distance. IMS-LD est un métalangage permettant de décrire des scénarios pédagogiques exprimant plusieurs types de pédagogies.

IMS Content Packaging permet d'empaqueter des scénarios IMS-LD dans un content package afin de les diffuser.

Nous venons de présenter des outils de transcription, des méthodes d'analyse et des spécifications de métadonnées s'appliquant aux corpus linguistiques et aux corpus de travail collaboratif. La partie suivante décrit les corpus dont nous disposons ainsi que la manière dont nous pouvons ou non utiliser les spécifications vues précédemment pour structurer nos corpus.

4. Discussion

Les spécifications étudiées permettent de décrire un corpus selon des points de vue différents. La TEI ainsi que d'autres spécifications permettent de décrire des corpus textuels et non audio ou vidéo. Cependant, nous souhaitons décrire les interactions dans des corpus multimodaux. IMS-LD permet de décrire le scénario pédagogique prescrit tandis que IMS-CP propose un formalisme pour le relier aux ressources utiles à l'exécution du scénario. Or, nous avons besoin de décrire, aussi, les productions des acteurs une fois la situation jouée. En effet, tant en vue d'une diffusion ciblée [PINCEMIN 99] de corpus linguistique que pour une diffusion pluridisciplinaire [BERNARD & GENDROT 05] de ce type de corpus, il est nécessaire de définir un format de structuration des corpus pour en favoriser l'accessibilité et la réutilisabilité. Cette nécessité est également mise en avant dans [FAERBER 05] qui souligne, en plus, la difficulté de définir des critères de structuration des documents répondant aux attentes de plusieurs disciplines.

5. Conclusion

L'étude des différents outils rend compte de l'absence d'un format d'échange de données consensuel entre les disciplines et pour des données de type hétérogène. De plus, les spécifications existantes ne répondent pas, en l'état, aux objectifs que nous nous sommes fixés en vue de décrire nos corpus hétérogènes : ouverture aux données d'un format hétérogène et des interactions multimédias, exhaustivité des données incluant les données du scénario joué. Nous inspirerons de ces spécifications pour les étendre à nos corpus, plus spécifiques, de formation en ligne.

Au vue de l'étude des outils, méthodes et spécifications existantes, nous souhaitons définir une spécification pour toutes les données des corpus et pour toutes les disciplines intéressées par ces corpus. Il s'agit d'un premier pas vers l'échange de corpus de formation en ligne. A terme, notre objectif de recherche est de concevoir un « socle/outil/plateforme » ouvert pour (1) permettre la mise à disposition des corpus respectant la spécification définie et (2) offrir des outils de transcription et d'aide à l'exploitation des données. Grâce à cela, la confrontation des méthodes et outils d'analyses mais également des résultats serait possible. Pour mener à bien ce travail de structuration, nous disposons de deux corpus recueillis exhaustivement lors de formations à distance médiatisées par une plateforme de formation à distance asynchrone, pour la première, et par une plateforme audio-

graphique synchrone pour la seconde. Ces formations avaient pour but respectif l'apprentissage du français en tant que langue étrangère pour un public anglophone et l'apprentissage de l'anglais sur objectifs spécifiques pour un public francophone.

6. Bibliographie

[AVOURIS et al. 02] Avouris N., Dimtracopoulou A., Komis V., Fidas C., « OCAF: An object-oriented model of analysis of collaborative problem solving », Proceedings CSCL 2002, Colorado, January 2002, G. Stahl (ed), p. 92-101.

[AVOURIS et al. 03] Avouris N., Komis V., Fiotakis G., Margaritis M., Tselios N., « A tool to support interaction and collaboration analysis of learning activities », *ICT 2003*, Tahiti, Papeete, French Polynesia, March 2003, p. 215-225.

[BENARD & GENDROT 05] Bénard F., Gendrot C., « Propositions de normalisation pour une base de corpus multimédia à l'ED268 », *RJC-ED268*, Université de la Sorbonne Nouvelle, Paris III, mai 2005.

[DE LA PASSARDIERE & JARRAUD 05] De La Passardière B., Jarraud P., « LOM et l'indexation de ressources scientifiques », *EIAH 2005,* Montpellier, 25-27 mai 2005, p. 57-68.

[FAERBER 05] Faerber R., « Indexer des situations d'apprentissage coopératif », *EIAH 2005,* Montpellier, 25-27 mai 2005, p. 321-332.

[GARG et al. 04] Garg S., Martinovski B., Robinson S., Stephan J., Tetreault J., Traum D.R., « Evaluation of transcription and annotation tools for a multi-modal, multi-party dialogue corpus », *4th International conference on language resources and evaluation (LREC 2004)*, Lisbon, Portugal, May 2004.

[HEID & MENGEL 99] Heid U., Mengel A., « Query language for research in phonetics », *ICPhS 99 (International Congress of Phonetic Sciences)*, San Francisco, August 1999.

[HENRI & CHARLIER 05] Henri F., Charlier B., « L'analyse des forums de discussion pour sortir de l'impasse », in G.L. Baron, E. Bruillard, and M. Sidir (Dir.), editors, *Symposium, formation et nouveaux instruments de communication*, Amiens, France, January 2005.

[IDE & ROMARY 02] Ide N., Romary L., « Standards for language resources », *LREC 2002*, Las Palmas, May 2002.

[KOPER & TATTERSALL 05] Koper R., Tattersall C., « Learning Design: A Handbook on Modelling and Delivering Networked Education and Training », Springer: Berlin Heidelburg New York, *Journal of Interactive Media on Education*, Special issue on « Learning Design », September 2005.

[PINCEMIN 99] Pincemin B., « Construire et utiliser un corpus: le point de vue d'une sémantique textuelle interprétative », *Atelier Corpus et TAL : pour une réflexion méthodologique, Conférence TALN'99*, Cargèse, 12-17 juillet 1999, Actes publiés par A. Condamines, M.-P. Péry-Woodley et C. Fabre, p. 26-36.

7. Références sur le WEB

[ISGS 06] Conférence ISGS 2005
http://vislab.cs.vt.edu/~gesture/multimodal/workshop/index.php, consulté en avril 2006
[TEI 06] TEI, http://www.tei-c.org/, consulté en avril 2006
[TC37/SC4 06] Comité ISO TC37/SC4, http://www.tc37sc4.org/, consulté en avril 2006