Constitution de Corpus d'Apprentissage et étiquetage simple sur des blogues interculturels

Mario Laurent, Thierry Chanier

Soumission pour l'atelier Traitement Automatique des Langues et EIAH : objets de recherche et objets d'étude. Assises du GDR I3 - mai 2012.

Constitution de Corpus d'Apprentissage et étiquetage simple sur des blogues interculturels

Laurent, M. & Chanier, T.

LRL, Université Blaise Pascal, Clermont-Ferrand.

http://lrl.univ-bpclermont.fr/

Mots-clés correspondant aux thèmes de l'atelier :

TAL et Traitement des traces pour le partage de corpus d'apprentissage, EIAH exploitant des outils de communication textuelle : blogues.

1. Corpus d'apprentissage MULCE

Nos travaux de recherche en EIAH nécessitent l'analyse de formations en ligne. Plus précisément, pour une formation donnée, nous souhaitons savoir si elle a été efficace, si les apprenants ont interagi et comment ils l'ont fait. Pour répondre à ces questions, nous travaillons après coup, à partir des traces recueillies pendant le déroulement de la formation.

Le recueil des traces est primordial, afin d'analyser une formation il est indispensable d'avoir accès à toutes les données qui lui sont liées. Or, du fait de l'utilisation de différents supports au cours d'une formation en ligne, les données sont souvent éparpillées. Elles peuvent se trouver en ligne dans des forums ou des blogues dont l'accès n'est pas garanti à long terme ou encore dans des formats propriétaires, spécifiques aux logiciels utilisés lors des échanges entre les apprenants, qui ne permettent pas de les réutiliser. Il est donc préalablement nécessaire de les extraire, de les réunir et de les organiser.

Mulce (MUltimodal contextualized Learner Corpus Exchange) a été créé afin de recueillir les données complètes de formations en ligne et de les organiser au sein de structures appelées corpus d'apprentissage ou LETEC (LEarning and TEaching Corpus) (Mulce, 2012a).

> *Un corpus d'apprentissage (LETEC) assemble de façon systématique et structurée un ensemble de données, particulièrement d'interactions, et de traces issues d'une expérimentation de formation partiellement ou totalement en ligne, enrichies par des informations techniques, humaines, pédagogiques et scientifiques permettant leur analyse en contexte.*

Chaque LETEC constitué est stocké dans la banque de corpus Mulce et mis à disposition en libre accès (Mulce, 2012b). La banque rassemble à ce jour 37 corpus sur 8 formations. Ces formations ont toutes eu lieu en ligne sur différents supports : forums, blogues, clavardages, plateformes audio-synchrones, etc... Les données associées sont multimodales : textuelles, audio et vidéo. Le tout représente des centaines de fichiers issus des différentes formations.

Pour un LETEC, toutes les informations sont rassemblées au sein d'une seule structure : Mulce-struct (2010). Il s'agit d'une structure XML étendue qui contient à la fois les métadonnées de la formation et le contenu des interactions, que celles-ci proviennent de blogues, de clavardages ou soient des transcriptions audio. Les métadonnées décrivent aussi bien les données contenues dans ce fichier XML, appelé manifeste, que les

ressources extérieures comme le scénario pédagogique de la formation, les consignes, les pré- et post-questionnaires et l'enregistrement des sessions audio ou vidéo. L'ensemble de ces fichiers est empaqueté dans une archive suivant le schéma standard IMS-CP (Chanier, 2010).

Cette démarche de constitution de corpus d'apprentissage permet à la fois la visibilité des travaux des chercheurs à l'origine des formations desquelles les données ont été extraites mais aussi la reproductibilité de leurs analyses. Chaque LETEC peut aussi être utilisé par de nouveaux chercheurs pour répondre à leurs propres problématiques et leurs résultats pourront venir enrichir la banque de données (Reffay et al., 2008). En plus des corpus d'apprentissage, d'autres objets sont présents sur la banque de corpus Mulce : les corpus distinguables. Ces corpus sont issus des corpus d'apprentissage mais n'en reprennent qu'une sous partie pour se concentrer sur l'analyse d'un phénomène précis (Chanier & Ciekansky, 2010).

2. Le cas d'INFRAL et de ses blogues

La formation Infral est l'objet d'étude d'un des LETEC. Infral est une formation au cours de laquelle des étudiants, futurs enseignants de français langue étrangère, de deux institutions (une université française et une allemande) ont échangé autour de thématiques interculturelles. Les apprenants étaient répartis en groupes mixtes appelés quadrems (deux étudiants de France et deux d'Allemagne dans chaque groupe).

Les échanges ont eu lieu en ligne, dans des blogues pour la partie asynchrone et dans une plateforme audio-synchrone pour la partie synchrone. L'un des objectifs du recueil des données d'Infral est d'analyser les interactions ayant eu cours dans ces blogues pour comprendre le rôle de chaque apprenant, leur participation et leur utilisation des différentes langues.

Aux enjeux habituels d'analyse linguistique s'ajoutent des phénomènes supplémentaires à étudier, propres à ces échanges. D'abord, la rédaction des messages sur un blogue amène à un type de discours particulier avec une segmentation particulière. Ensuite, deux langues (français et allemand, avec prédominance du français) sont utilisées de façon mixte dans les messages, parfois dans une même phrase. Enfin, les apprenants ont produit un nombre important de graphies (*token*) non standards, que ce soit du à des erreurs typographiques ou lexicales.

Pour constituer le corpus d'apprentissage Infral, il a fallu récupérer le contenu des blogues, exportable seulement au format ATOM, puis le transformer au format exigé par Mulce-struct. Pour cela, nous avons procédé à un premier traitement automatique, en utilisant le langage de transformation XSLT.

Précédemment, l'équipe de recherche, qui avait monté l'expérimentation INFRAL, avait commencé à publier avant la constitution du corpus d'apprentissage (Abendroth-Timmer et al., 2009). Les premières analyses concernaient le rôle et la participation de chaque apprenant. Pour ce faire, les chercheurs avaient effectué des comptages des messages, des graphies et de l'utilisation plus ou moins importante d'une langue par rapport à l'autre, par quadrem et par apprenant. Ces calculs avaient été effectués en partie à la main, y compris une évaluation grossière des proportions de graphies appartenant à chaque langue.

Une fois le corpus Infral constitué, l'objectif était de vérifier, affiner ces calculs et commencer à développer des outils appropriés à ce type de traitement.

3. Programme Python et corpus distinguable

Pour résoudre le problème du comptage des graphies par langue, nous avons entrepris d'effectuer des traitements automatiques avec le langage Python et la bibliothèque de fonctions NLTK (Natural Language ToolKit) sur les messages et les commentaires d'Infral (Laurent, 2011a). Nous nous sommes orientés vers l'utilisation de NLTK (2012) parce que nous n'avions pas trouvé, auparavant, d'étude linguistique concernant les blogues ni d'outils de TAL destinés à traiter ce genre particulier.

Le programme *Python* que nous avons développé permet, à partir du texte brut, de segmenter chaque message en graphie et d'attribuer à chaque graphie une étiquette de forme (*type*) et une étiquette de langue. Le tout ressort dans un fichier XML structuré, respectant la syntaxe, les éléments et les attributs de la TEI, que nous avons ensuite complété par des métadonnées (TEI, 2012).

Voyons comment fonctionne le programme de manière détaillée. Tout d'abord, le texte passe par une fonction de segmentation. Cette fonction est basée sur une expression régulière *Python* et inspirée des fonctions de segmentation du livre *NLTK* (Bird, 2009). Une fois les messages segmentés, plusieurs solutions sont possibles pour leur attribuer une étiquette de forme et une étiquette de langue. Afin d'obtenir un taux minime d'erreur, l'idéal est d'utiliser un corpus de référence, le plus volumineux possible, qui est déjà étiqueté comme on le souhaite et d'entraîner un programme sur celui-ci. De tels programmes ont été développés par la communauté *NLTK*. Ils permettent de dégager statistiquement un ensemble de règles prenant en compte le contexte pour attribuer une étiquette à chaque graphie. Cette stratégie est appelée n-gram où n est le nombre d'éléments du contexte pris en compte. Du fait de l'absence de corpus similaire déjà constitué, nous avons opté pour une autre solution qui est d'étiqueter de façon semi-automatique.

Pour cela, le programme utilise des dictionnaires, français et allemand, inclus dans la bibliothèque *Pyenchant* (2012), ces dictionnaires sont ceux utilisés par *OpenOffice* (OpenOffice dictionnaries, 2012). Ils comprennent, pour chaque langue, l'ensemble des formes fléchies. Le programme parcourt ces dictionnaires pour chaque graphie et attribue simultanément les étiquettes de forme et de langue, lorsque la graphie courante est reconnue dans l'un des dictionnaires. Pour chaque graphie non-standard, donc absente du dictionnaire, rencontrée, l'utilisateur est interrogé pour déterminer la langue correspondante. Puis, une liste de propositions de formes, correspondant à la graphie, est proposée à l'utilisateur. Les propositions sont générées grâce à un appel à une autre fonction de *Pyenchant*. Le programme propose ensuite d'ajouter la graphie à un dictionnaire. Il laisse aussi à l'utilisateur la possibilité de garder en mémoire ses choix d'associations entre graphie et forme, pour éviter d'avoir à traiter de nombreuses fois la même graphie non standard et ainsi automatiser davantage cette étape.

Une fois les étiquettes de langue et de graphie connues, le fichier de sortie XML est généré en faisant appel au module *Python* « lxml.etree » (Lxml, 2012). Le fichier de sortie respecte les normes de la TEI.

Comme le montre l'extrait ci-dessous, chaque message est inclus dans une balise , spécifiant l'identifiant du message, son type et son éventuel lien avec un autre message. On trouve ensuite une balise <u> qui spécifie l'auteur du message. Enfin, le contenu du message est segmenté en graphie, chacune entre des balises <w>. Chaque balise <w> contient des attributs qui spécifient la forme associée et la langue correspondante. En TEI, l'attribut de langue est noté «xml:lang ». Pour l'attribut de forme, nous l'avons ajouté dans les déclarations de la feuille de style puisqu'il n'est pas standard en TEI.

```
<div postID="3983298308313693092" type="message">
```

```
            <u who="afbes1_2, Johanne">
[…]

              <w xml:lang="fr-FR" forme="Où">Où</w>
              <w xml:lang="fr-FR" forme="exactement">exactement</w>
              <w xml:lang="fr-FR" forme="car">car</w>
              <w xml:lang="fr-FR" forme="j'">j'</w>
              <w xml:lang="fr-FR" forme="habite">habite</w>
[…]

              <w xml:lang="de-DE" forme="Bald">Bald</w>
            </u>
          </div>
```

Le tout a été assemblé dans un corpus de type distinguable disponible sur le site de la banque Mulce (Laurent, 2011b). On y retrouve à la fois le fichier de sortie, les fichiers d'entrée et le programme lui-même, accompagné de la documentation sur le corpus. Pour de prochaines analyses ou une extension des étiquettes, nous pourrons désormais travailler directement à partir de ce corpus distinguable.

4. Perspectives

Par la suite, nous souhaitons associer des étiquettes de lemme à chaque graphie afin de pouvoir analyser la richesse lexicale des messages de chaque apprenant. Nous cherchons également à renseigner des étiquettes de catégorie grammaticale pour d'autres analyses linguistiques, que ce soit pour le corpus d'apprentissage de la formation Infral ou les autres corpus de la banque de données Mulce. Différentes pistes s'offrent à nous :

- Utiliser NLTK pour les catégories grammaticales. Ceci n'est possible qu'en utilisant une combinaison de *n-gram taggers* et en entraînant un programme à l'étiquetage. Ces programmes existent déjà, il suffirait donc, pour les faire tourner et obtenir de bons résultats, d'avoir un corpus de référence, en français, sur des échanges en ligne.

- Adapter les fonctionnalités développées sur CLAN (Parisse & Le Normand, 2000) pour l'étiquetage de la langue parlée au traitement de la langue écrite, comprenant des graphies non standards.

5. Références

Tous les liens Internet de cette section ont été vérifiés en date du 13 avril 2012.

Abendroth-Timmer, D., Bechtel, M., Chanier, T. & Ciekanski, M. (2009). "From developing to investigating intercultural competence in practice through oral and written interactions in online exchanges", Kongress für Fremdsprachendidaktik der Deutschen Gesellschaft für Fremdsprachenforschung (DGFF-Tagung), Universität Leipzig, octobre 2009. [http://edutice.archives-ouvertes.fr/edutice-00548891/]

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media. [http://www.nltk.org/book]

Chanier, T. (2010). Structure des objets de l'archive Mulce, *Mulce.org Documentation* [Site internet]. [http://mulce-doc.univ-bpclermont.fr/spip.php?article25]

Chanier, T. & Ciekansky, M. (2010). Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage, *Apprentissage des Langues et Systèmes d'Information et de Communication (Alsic), 13*. [http://alsic.revues.org/1666]

Laurent, M. (2011a). Structuration des données des blogues de la formation Infral à l'aide des outils de programmation Python et NLTK. Mémoire de Master 2 Sciences du Langage, Université Blaise Pascal.

Laurent, M. (2011b). Blog's data from Infral structured, tokenized and tagged into a XML file, *banque de corpus Mulce* [site internet]. [http://mulce.univ-bpclermont.fr:8080/PlateFormeMulce/VIEW/PUBLIC/03/VMeta.do?adr=Infral%2FCorpus_objets%2Fmce-infral-tagged_blogs]

Lxml (2012). Site de documentation de la bibliothèque de fonctions lxml, pour le langage Python [site internet]. [http://lxml.de/]

Mulce (2012a). Site de documentation du projet Multimodal Learning Corpus Exchange (2007-2012) [site internet]. [http://mulce.org]

Mulce (2012b). Site de la banque de corpus Mulce [site internet]. Université Blaise Pascal [http://mulce.univ-bpclermont.fr:8080/PlateFormeMulce/]

Mulce-struct (2010). Schéma de la structure d'un corpus LETEC Mulce. [http://lrl-diffusion.univ-bpclermont.fr/mulce/metadata/mce-schemas/]

NLTK (2012). Ensemble de modules Python pour le NLP, documentation, données et analyses [site internet]. [http://www.nltk.org/]

OpenOffice dictionnaries (2012). Documentation et liens de téléchargements libres vers les dictionnaires utilisés par la suite OpenOffice 2.x [site internet]. [http://wiki.services.openoffice.org/wiki/Dictionaries]

Parisse C. & Le Normand M.-T. (2000). Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research, Methods, Instruments and Computers,* 32 (3). pp. 468-481

Pyenchant (2012). Site de documentation de la bibliothèque de fonctions Pyenchant, pour le langage Python [site internet]. [http://packages.python.org/pyenchant/tutorial.html]

Reffay, C, Chanier, T., Noras, M. & Betbeder, M.-L. (2008). Contribution à la structuration de corpus d'apprentissage pour un meilleur partage en recherche, in Basque, J. & Reffay, C. (dir.), *numéro spécial EPAL (échanger pour apprendre en ligne), Sciences et Technologies de l'Information et de la Communication pour l'Education et la Formation (STICEF),* 15. [http://sticef.univ-lemans.fr/num/vol2008/01-reffay/sticef_2008_reffay_01.htm]

TEI (2012). Principes directeurs pour l'encodage et l'échange de textes électroniques [site internet]. [http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/index.html]

****

Mario Laurent est doctorant au Laboratoire de Recherche sur le Langage (LRL) depuis 2012. Il a obtenu en 2011 son Master recherche en sciences du langage à l'Université Blaise Pascal de Clermont-Ferrand. Son sujet de thèse est lié au TAL et à l'aide à l'apprentissage et s'intitule *Recherche et développement d'un Logiciel Intelligent de Cartographie Inversée (LICI) pour l'aide à la compréhension de texte par des publics sujets à des troubles spécifiques du langage*.

Contact:

Mario Laurent

mario.laurent@etudiant.univ-bpclermont.fr

http://lrl.univ-bpclermont.fr/spip.php?article325

Thierry Chanier

thierry.chanier@univ-bpclermont.fr , Université Blaise Pascal

http://lrl.univ-bpclermont.fr/spip.php?rubrique98

LRL, Maison des Sciences de l'Homme

4 rue Ledru - 63057 Clermont-Ferrand Cedex 1