Vers une utilisation du TAL dans la description pédagogique de textes dans l'enseignement des langues

Mathieu LOISEAU

LIDILEM – Université Stendhal Grenoble 3
mathieu.loiseau@u-grenoble3.fr

Mots-clefs – Keywords

Corpus, ALAO, TAL, indexation pédagogique, ressources textuelles

Corpus, CALL, NLP, pedagogical indexation, textual resources

Résumé – Abstract

Alors que de nombreux travaux portent actuellement sur la linguistique de corpus, l'utilisation de textes authentiques en classe de langue, ou de corpus dans l'enseignement des langues (via concordanciers), quasiment aucun travail n'a été réalisé en vue de la réalisation de bases de textes à l'usage des enseignants de langue, indexées en fonction de critères relevant de la problématique de la didactique des langues.
Dans le cadre de cet article, nous proposons de préciser cette notion d'indexation pédagogique, puis de présenter les principaux standards de description de ressources pédagogiques existants, avant de montrer l'inadéquation de ces standards à la description de textes dans l'optique de leur utilisation dans l'enseignement des langues. Enfin nous en aborderons les conséquences relativement à la réalisation de la base.

Despite numerous works concerning corpus linguistics, the use of authentic texts in language teaching as well as corpora in language teaching (through concordancers), hardly any work deals with the creation of a teacher-friendly text base that would be indexed according to criteria relevant to the set of problems of language didactics.
In this article, we will first discuss the notion of pedagogical indexation. We will then present the principal pedagogical resource description standards, before showing that those standards are inadequate to handle our problem. Finally we shall give an overview of the consequences of these inadequacies on the implementation of the text base.

1 Indexation pédagogique pour l'enseignement des langues

Parmi les différentes plateformes d'apprentissage des langues assisté par ordinateur (ALAO), la plateforme MIRTO, en cours de développement à l'université Stendhal – Grenoble 3, se

démarque grâce à une approche résolument centrée sur la didactique des langues. Alors que de telles plateformes forcent, en règle générale, l'utilisateur enseignant – didacticien à reformuler sa problématique en termes informatiques, MIRTO reste avant tout un produit didactique : « *un programme qui met en oeuvre une solution didactique pour un problème de la didactique des langues sans altérer, ni la solution, ni, a fortiori, le problème.* » (Antoniadis et al., 2004). Ainsi MIRTO permet la conception, sans compétences informatiques préalables, d'activités didactiques. A l'heure actuelle, MIRTO est capable, à la donnée d'un texte brut et d'une série de paramètres entrée par l'enseignant, de générer divers types d'activités parmi lesquels des exercices lacunaires ou des exercices de traduction avec aide…

Cette philosophie est à l'origine du projet de création d'une base de textes indexée pédagogiquement pour l'enseignement des langues. En effet, grâce à son architecture, MIRTO est capable, de générer autant d'activités d'un type donné que l'on sera capable de lui fournir de textes. D'où l'idée d'intégrer un corpus à la plateforme. Cependant, deux textes différents ne sont pas strictement interchangeables pour une activité donnée. Les propriétés de chaque texte les rendront plus ou moins adaptés à une activité donnée. Un tel corpus ne pourra donc pas être une simple collection de textes, cette collection devra être organisée, indexée.

Depuis le lancement de ce projet, sa portée a été élargie, non seulement la base de textes devra être intégrée à la plateforme MIRTO, mais aussi pouvoir exister indépendamment de celle-ci et permettre à des enseignants de la consulter pour trouver des textes à utiliser en classe.

L'approche choisie pour la conception de cette base, s'inscrit dans la philosophie MIRTO, dans la mesure où ses trois fonctionnalités clés seront centrée sur l'enseignant :

- Exécution de requêtes selon des critères relevant de la didactique des langues.
- Ajout de textes dans la base, un processus qui doit être automatisé autant que faire se peut, pour ne pas rendre cette opération dissuasive de part sa complexité ou sa durée.
- Présence d'une interface qui permette aux enseignants n'ayant pas de connaissances particulières en informatique de pouvoir utiliser la base.

Ces fonctionnalités reposeront sur ce que l'on appellera l'indexation pédagogique de la base. On dira que des objets ont été indexés pédagogiquement, s'ils ont été indexés selon un système les décrivant en fonction de critères pédagogiques (relevant de la problématique de la didactique). Ici, les objets sont des textes et les critères pédagogiques relèvent de la didactique des langues. On parlera donc d'indexation pédagogique pour l'enseignement des langues.

Le travail d'indexation pédagogique en soi, reviendra donc, a priori, d'une part, aux utilisateurs du système et d'autre part au système lui-même, puisque lors de l'ajout d'un document dans la base, une partie du traitement sera vraisemblablement automatisée. C'est donc la définition du langage documentaire qui nous incombe. Elle repose sur trois problèmes distincts et interdépendants :

- L'utilisation des textes dans l'enseignement des langues, que ce soit avec ou sans l'intervention de l'ordinateur
- Le processus de recherche d'un texte par un enseignant de langue que nous modéliserons, à travers le recensement de critères de recherche.
- L'implémentation informatique de la base.

Nous détaillerons ici principalement une partie du troisième de ces aspects, même si aucun des trois aspects ne peut être traité complètement indépendamment des deux autres. Pour ce faire, nous utiliserons les résultats d'une étude préliminaire, qui nous a permis de commencer à mettre en évidence certains besoins des enseignants, comme le fait de pouvoir choisir un texte en fonction de son contenu lexical et grammatical ou encore celui d'avoir recours à des textes entiers et non à des portions de texte. Les résultats de cette étude ne sont pas définitifs, ils ont servi de base à la création d'un questionnaire à plus grande échelle. Ils permettent cependant d'évaluer l'adéquation des systèmes existants avec la notion d'indexation pédagogique de textes pour l'enseignement des langues.

Avant de proposer une solution *ad hoc* pour le langage documentaire à utiliser, nous devions nous assurer que l'existant ne répondait pas d'ores et déjà à notre problème, nous présenterons donc tout d'abord les principaux standards de description de ressources pédagogiques, après quoi nous les soumettrons à notre problématique avant de conclure sur le sujet.

2 Standards de description de ressources pédagogiques

2.1 Présentations des principaux standards

La Dublin Core Metadata Initiative (DCMI) constitue le standard le plus ancien parmi ceux que nous allons présenter. Bien que n'étant pas à proprement parler un standard de description « pédagogique », il a influencé les principaux standards de description d'objets pédagogiques existants, d'où sa présence ici. La DCMI est à l'origine du Dublin Core Element Set (DCES) qui permet de décrire des ressources à partir de quinze éléments, tous utilisables zéro, une ou plusieurs fois (Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights) (DCMI, 2005). Ces éléments très généraux ne permettront pas de décrire suffisamment précisément tout type de document mais le DCES n'est pas figé :

Ce modèle permet aux différentes communautés d'utilisateurs de se servir des éléments DC pour les informations descriptives fondamentales tout en autorisant les extensions spécifiques à un domaine et qui sont pertinentes pour un public moins large (Hillman, 2005) (Traduction de l'auteur)

Le Gateway to Educational Material (GEM, 2004) et Educational Network of Australia metadata (EdNA, 2002) constituent tous deux des exemples d'extensions / raffinements du DCES pour la description d'objets pédagogiques.

Enfin, le LOM (Learning Object Metadata), qui est probablement le plus utilisé et le plus influent de ces formats, est le fruit de la collaboration d'équipes du projet européen ARIADNE[1] et du consortium américain IMS Global Learning[1].

[1] http://www.ariadne-eu.org et http://www.imsproject.org respectivement

2.2 Caractéristiques communes

Tous ces standards restent très généralistes. C'est évident pour la DCMI, cependant, c'est le cas aussi pour les autres. Ils se veulent capable de décrire des ressources appartenant à n'importe quel domaine de l'enseignement : dans le LOM un objet pédagogique est « *N'importe quelle entité, numérisée ou non, qui pourrait être utilisée pour l'apprentissage, l'enseignement ou la formation* » (LOM, 2002) (traduction de l'auteur)

LOM utilise des niveaux d'agrégation censés permettre à partir du même ensemble d'éléments de décrire aussi bien un texte qu'une formation entière. Le fait que ces standards soient si généralistes nuira fatalement à la précision de la description qu'elles proposent, comme le fait remarquer Jean-Philippe Pernin lorsqu'il cite parmi les imprécisions ou ambiguïtés dont souffre le LOM : « *la volonté d'intégrer au sein d'un même modèle des entités de niveau conceptuellement très différent* » (Pernin, 2004). Notre problème est extrêmement spécifique : la description que l'on pourra faire d'un texte dans le cadre de l'enseignement des langues, ne sera pas nécessairement applicable à la description d'un texte dans une autre matière et ne le sera assurément pas pour décrire un problème de mathématique ou un cursus entier.

Les standards présentés sont donc trop généralistes pour être appliqués tels quels dans notre projet, mais elles présentent toutes la possibilité d'être adaptées à d'autres problématiques. La conformité au LOM s'exprime en les termes suivants (LOM, 2002) :

- il n'est pas obligatoire de renseigner tous les éléments LOM
- si l'on n'étend pas le standard l'instance sera strictement conforme
- sinon elle sera conforme à condition qu'aucun des éléments ajoutés ne remplace un élément LOM

On pourra donc tout à fait réutiliser le LOM, on sortira du cadre de la stricte conformité mais cela ne l'exclut pas pour autant. EdNA Metadata et GEM, constituant des extensions du DCES, ils pourront eux aussi être raffinés en respectant les principes de la grammaire DCMI.

Que ce soit dans LOM, le GEM ou EdNA metadata, les éléments « pédagogiques » sont séparés des autres éléments. Le GEM et EdNA suivent les directives DCMI et utilisent donc le DCES pour « *les informations descriptives fondamentales* ». Dans LOM, les objets pédagogiques sont décrits en 77 éléments répartis en 9 catégories dont l'une et dédiée à la composante pédagogique de la description (« Educational »).

2.3 Les éléments pédagogiques

Le GEM ajoute cinq éléments dits pédagogique (Audience, Duration, Essential Ressources, Instructional Method et Standards) (GEM, 2004), EdNa Metadata trois (Audience, Category Code et Review) (EdNA, 2002) et enfin le LOM dispose de onze éléments de ce type (5.1. Interactivity Type, 5.2. Learning Resource Type, 5.3. Interactvity Level, 5.4. Semantic Density, 5.5. Intended End User Role, 5.6. Context, 5.7. Typical Age Range, 5.8. Difficulty, 5.9. Typical Learning Time, 5.10. Description, 5.11. Language).

Quelle que soit le standard utilisé, chacun des champs concerne une propriété jugée intrinsèque à l'objet pédagogique. Pour le GEM, EdNA Metadata et LOM, un texte brut entre

dans le cadre de l'utilisation du standard. Pourtant les propriétés considérées comme intrinsèques à un objet pédagogique ne le sont pas forcément pour un texte brut.

Prenons l'exemple des éléments « *Audience* » de GEM et EdNA, qui pourraient correspondre à l'élément 5.5 ou au couple d'éléments 5.5 et 5.7 de LOM. Il est ressorti de l'étude préliminaire évoquée précédemment que selon le type d'activité, un texte donné pouvait être utilisé avec des publics très variés. Pour une activité de compréhension, il est moins important que les apprenants connaissent parfaitement les structures et le vocabulaire employés que dans un exercice de grammaire. Le texte ne peut donc pas être considéré comme intrinsèquement destiné à un public puisqu'en fonction du type d'activité, le public pourra être différent. De même les éléments « *Duration* » du GEM et 5.9 de LOM, que l'on peut considérer comme équivalents, n'auront aucun sens pour un texte brut. Un texte donné pouvant servir de support à un exercice de compréhension ou à l'introduction d'une nouvelle notion grammaticale sera utilisé beaucoup plus longtemps que le même texte servant uniquement de base pour un exercice lacunaire concernant les déterminants pour un public plus avancé. Nous terminerons ce tour d'horizon non exhaustif des champs pédagogiques en faisant remarquer qu'un champ comme l'élément 5.3 de LOM n'est pas seulement inadapté à notre problème, il est en outre complexe à renseigner puisque les valeurs acceptées sont à choisir parmi « *very low, low, medium, high et very high* » sans que le standard ne fournisse de réelles directives pour leur utilisation.

3 Conclusion

Mêle si ces différents standards, ne sont pas utilisables tels quels pour notre projet, il n'est pas exclu de réaliser un profil d'application de métadonnées : « *un assemblage d'éléments de métadonnées choisis à partir d'un ou plusieurs schémas de métadonnées et combinés dans un schéma composite* ». (Duval et al., 2002) (traduction de l'auteur) En effet la plupart des éléments non pédagogiques seront réutilisables. De plus, la création d'un profil d'application de métadonnées peut permettre d'améliorer les ensembles de valeurs proposés pour certains éléments, en fournissant des lignes directrices très précises quant à l'utilisation d'une norme.

Les remarques que nous avons pu faire sur le caractère intrinsèque ou non de certaines propriétés décrites dans les éléments pédagogiques nous ont amené à préciser l'architecture probable du système de gestion de la base de texte. Le fait que les éléments ne correspondent pas à des propriétés intrinsèques des textes ne signifie pas que certaines informations qu'ils pourraient contenir ne s'avéreront pas intéressantes pour les enseignants. La description des textes devra, pour être cohérente, contenir exclusivement des informations valables quelle que soit l'utilisation du texte. Nous nous proposons donc de séparer le processus en deux parties. Dans un premier temps on décrira les textes en fonction de traits pertinents pour les enseignants, mais qui ne varieront pas selon l'utilisation qui sera faite du texte. Les caractéristiques lexicales, syntaxiques (vocabulaire, structures grammaticales) du texte, par exemple sont des caractéristiques intrinsèques au texte, qui pourront dans une certaine mesure être relevées automatiquement grâce à l'utilisation d'outils TAL comme un lemmatiseur ou un analyseur morphologique. Mais les propriétés intrinsèques aux textes ne seront probablement pas suffisantes pour répondre à toutes les questions des utilisateurs de la base. Nous aurons donc recours, à une couche logicielle supplémentaire, un moteur d'inférence qui permettra de considérer plusieurs facettes de chaque texte. Le moteur devra analyser les caractéristiques de chaque texte en les combinant avec des informations fournies par

l'utilisateur dans sa requête afin d'en déduire certaines qualités d'un texte en fonction de l'utilisation qui en sera faite.

Il nous est actuellement impossible de faire la liste des outils TAL nécessaires à la réalisation de cette base. Les deux exemples ci-dessus (lemmatiseur / analyseur morphologique) paraissent correspondre à la demande des enseignants[2]. Cependant cela devra être confirmé par une analyse plus poussée des besoins des enseignants, via un questionnaire moins ouvert. En précisant les besoins, nous serons en mesure de dire si oui ou non ces outils peuvent répondre à la demande des enseignants. En outre, les résultats obtenus avec ces types d'outils sont suffisamment fiables pour que ces derniers puissent être employés. Car si l'utilité de l'outil dans la perspective de la description du texte pour l'enseignement des langues est une valeur primordiale pour décider de son utilisation, elle n'est pas la seule, une fiabilité minimum sera à définir plus tard afin que la précision des requêtes effectuées sur la base ne soit pas trop faible. Enfin, nous devrons probablement adopter une architecture modulaire pour le système, afin de pouvoir intégrer plus tard de nouveaux outils à la base.

Références

Antoniadis, G., Echinard, S., Kraif, O., Lebarbé, T., Ponton, C. (2004), Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO 57-70 actes de la journée TAL et Apprentissage des langues du 22 Octobre 2004, http://www.u-grenoble3.fr/lidilem/talal/actes/JourneeTALAL-041022-antoniadis.pdf.

DCMI Usage Board (10 janvier 2005). DCMI Metadata Terms. Consulté en janvier 2005 http://dublincore.org/documents/dcmi-terms/

Duval E., Hodgins W., Sutton S., Weibel S. L. (2002) Metadata Principles and Practicalities. D-Lib Magazine, 8 (4). Consulté en avril 2005. http://www.dlib.org/dlib/april02/weibel/04weibel.html

EdNA (septembre 2002), EdNA Metadata Standard V1.1. Consulté en septembre 2004 http://www.edna.edu.au/edna/go/pid/385

GEM (1er Juin 2004), GEM Top-Level Elements. Consulté en janvier 2005 http://www.thegateway.org/about/documentation/metadataElements/

Hillman D. (26 août 2003), Using Dublin Core Consulté en septembre 2004 http://dublincore.org/documents/usageguide/

LOM (juillet 2002), Final 1484.12.1 LOM Draft Standard Document. Consulté en avril 2005 http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Pernin J.P. (2004), A propos des objets pédagogiques, in "Entre technique et pédagogie : la création de contenus multimédia pour l'enseignement et la formation", Neuchâtel : IRDP.

[2] Données provenant d'un premier questionnaire, volontairement très ouvert de manière à laisser les enseignants s'exprimer sans qu'on leur impose un point de vue a priori, rempli par 130 enseignants.