Pedagogical text indexation and exploitation for language learning

Mathieu Loiseau, Georges Antoniadis, Claude Ponton

International Conference on Multimedia and ICT in Education. 22 - 24 April 2009, Lisbon, Portugal www.formatex.org/micte2009/

Pedagogical text indexation and exploitation for language teaching

M. Loiseau[*,1], **G. Antoniadis**[1], and **C. Ponton**[1]

[1] LIDILEM, Université Stendhal Grenoble 3, BP 25 - 38040 Grenoble cedex 9, France

In this article we present the MIRTO platform – under development at the University Stendhal of Grenoble – and how it addresses common flaws of CALL software. This platform led to another project: the creation of a pedagogically indexed text base. We introduce here the notion of pedagogical indexation, and confront the particular case of pedagogical indexation for language learning with the existing pedagogical resource description standards, before proposing leads towards the implementation of the former.

Keywords CALL, NLP, corpus, pedagogical indexation

1. MIRTO or integrating NLP in CALL

In Antoniadis et. al.[1], we identify current CALL software flaws: the poorness of meaning associated to any linguistic sequence, the rigidity of software and the necessity for language teacher users to express their pedagogical solutions in computer understandable terms instead of resorting to language didactics. These flaws mostly stem from the divergences between computer science's and didactics' view of the notion of "language". "*Computer science can only consider and process the form of language independently of any interpretation, while, for language didactics, the form only exists through its properties and the concepts it is supposed to represent*"[2].

The MIRTO platform, currently under development at the University Stendhal of Grenoble, plans to address these problems via the use of NLP (Natural Language Processing) tools and collaborative work with didactics experts. In order to support this approach, we resorted to the following architecture:

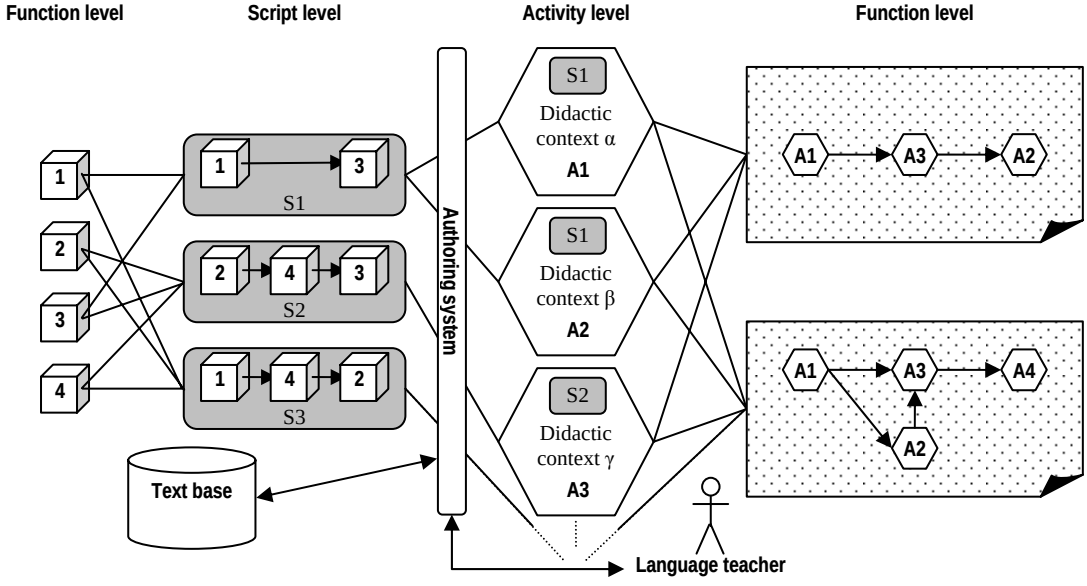

Fig. 1 the global architecture of MIRTO

[*] Corresponding author: e-mail: mathieu.loiseau@u-grenoble3.fr, Phone: +33 4 76 86 22 57

Functions are NLP tools providing basic NLP processes (e.g. tokenization). Due to their technical nature they are not accessible to the end user. These tools are grouped into scripts by computer engineers considering didactic requests made by teachers, thus providing pedagogically aimed toolboxes. The next two levels are the most explicitly didactics oriented. Activities are created by the user (language teachers) through the authoring system which offers a user friendly interface for the definition of the didactic context. Let us take the example of a gap filling exercise script. The creation of an activity will be done by specifying the didactic context consisting of the text (to which the script will be applied), the features of the words that ought to be removed (e.g. simple past conjugated verbs), the instructions that the learner will receive, the evaluation strategy that should be used and potentially other parameters such as the assistance to be provided to the learner. Finally the scenario level allows the definition of non linear sequence of activities. Intelligent error detection (and corresponding feedback) added to the development of the scenario level should improve the ability of the system to adapt the course of learners, while the modularity described above along with the user interface, seeks to provide a more pedagogically sound product. Finally the use of customizable scripts for the creation of activities seeks to make MIRTO as flexible as possible.

To further improve this flexibility, MIRTO will, in term, include a text base. The text base, in the context of MIRTO, offers an opportunity to separate data from structures (data being the texts and structures the scripts). Doing so enhances the possibilities for each type of activity. Each script along with a given set of parameters can be used with any given text in the appropriate language, thus creating a set of activities according to the same pattern. Still, despite the fact that technically any text/script combination is possible (provided the script is meant to work with the text's language), some combination might not be appropriate didactic-wise. Such a decision is ultimately the language teacher's to take. This decision can be aided by properly indexing the text base, that is to say indexing it pedagogically for language teaching.

2. Pedagogical indexation

2.1. Definition

Before going further in the description of the text base, we should properly define what we mean by pedagogical indexation. Indexation depends on the "documentary language" according to which it is performed. Lefèvre defines a documentary language as an "*artificial language, which provides a formalised and univocal representation of the documents of a corpus and of the questions interesting a group of users, so as to allow the simple spotting of the documents of the corpus which answer the questions of those users.*"[1][3] As this definition suggests it, the definition of a documentary language has to be done according to the potential users of the data base it will index. Therefore, pedagogical indexation will concern the indexation of objects, under the scope of their potential use by teachers in the course of their teaching. Hence the definition of pedagogical indexation: "*indexation performed following a documentary language describing the objects indexed according to pedagogical criteria (relevant to didactics)*"[4]. This definition is purposely very generic. Consequently our work is a particular case of pedagogical indexation, which we will address as pedagogical indexation for language teaching.

2.2. In and out of MIRTO

In the context of MIRTO, indexing pedagogically for language teaching a text base will not only allow teachers to have easier and quicker access to a set of texts satisfying their queries, from which to select the most appropriate to their needs. It might also allow the definition of a set of required features for

[1] "Langage artificiel qui fournit une représentation formalisée et univoque des documents d'un corpus et des questions qui intéressent un groupe d'usagers, afin de permettre le repérage simple des documents du corpus qui répondent aux questions de ces usagers." (translated by the author)

texts in order to consider them usable in a certain activity. This would allow learners to repeat a same activity without answering the same question over and over.
We thought it reductive to limit the text base to its use in MIRTO. Even though, it stemmed from the latter, they now constitute two different projects. The integration of the text base in the platform is a perspective but we consider it as a stand-alone aid for the teachers in their preparation of any type of class activity.

3. Pedagogically indexed text base (for language teaching)

The text base will have to feature the following functionalities:

- Allow teachers to perform queries according to criteria consistent with language didactics.
- Allow teachers to add his/her own text to the text base.
- Language teachers are not necessarily computer savvy; the interface has therefore to be as user friendly as possible.

This project is resolutely user centered. This means that we will try to automate as much as we can the process of addition of a new text to the database, by using computer tools (NLP or not). This also means that the definition of the documentary language for the text base must be performed in close collaboration with teachers. We started by performing a preliminary study based on a series of interviews with language teachers of various backgrounds and experience. This study has allowed us to have a better idea of how teachers use texts in their class and how they look for them. To generalize the results of this preliminary study we set up a pair of questionnaires, the first of which has already been put online and filled in by over 130 teachers. Those studies are meant to allow us to develop the documentary language for our text base.

4. Pedagogical resource description standards

Not much has been done towards the implementation of pedagogical text bases; we thus chose to study the existent, that is to say the main pedagogical resource description standards, hoping that they can be re-used in our project. We mainly examined the Dublin Core Metadata Element Set (DCES), Learning Object Metadata (LOM), the Getaway to Educational Material metadata (GEM) and the Educational Network of Australia metadata (EdNA). We based our reflection on the answers we got from our preliminary study and first questionnaire.

4.1. Presentation of the standards

The first of those standards to be created was the DCES[5] as part of the Dublin Core Metadata Initiative. It does not strictly concern pedagogical resources, but occupies a central position among the other standards. GEM[6] and EdNA[7] metadata are indeed extensions of the DCES. LOM is not technically an extension of DCES, but LOM specifications acknowledge the influence of the Dublin Core in the creation of the standard. There even exists a memorandum of understanding between the two organizations[2].
Despite their differences, those standards share some features: the fact that they are metadata element sets, the fact that they all stem from DCES. In the rest of the article we will mostly describe LOM and specify the analogies, which exist between the diverse standards.

4.2. Learning Object Metadata

LOM is probably the most used standard among those presented; it is used in SCORM for instance. It defines a set of data elements grouped into nine categories used to describe learning objects. None of the data elements are mandatory to maintain conformance. In LOM, a learning object is "*any entity – digital*

[2] http://dublincore.org/documents/dcmi-ieee-mou/index.shtml

or non-digital – that may be used for learning, education or training"[8]. LOM specifies four aggregation levels, which further qualify the objects and emphasize the extensiveness of the definition above:

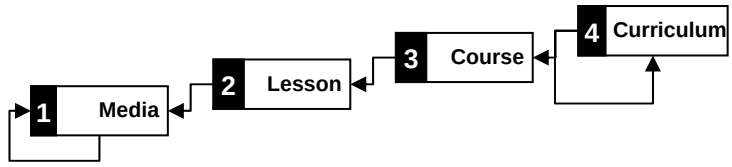

Fig. 2 LOM aggregation model The number is the aggregation level and the item in the box is an example of an object of such aggregation level. The arrows are to be read: "can be composed of".

A text is therefore an aggregation level 1 learning object and fits in with LOM (as it does with GEM and EdNA). Now that the applicability of LOM to our project has been established, we are going to examine the data elements of the educational category.

4.3. Pedagogical data elements

There are 11 data elements in the LOM category labeled "5. Educational": 5.1. Interactivity Type, 5.2. Learning Resource Type, 5.3. Interactivity 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. These elements as well as all the other elements of the standards presented here concern characteristics considered inherent to the object. This proves to be a problem as far as pedagogical properties are concerned.

Let us take the example of the data elements "Intended End User Role" and "Typical Age Range" of LOM, which roughly correspond to the "Audience" elements of GEM and EdNA. In an activity, one can consider the role of the learner as an inherent characteristic but in the case of a text, free from any didactic context, it is not the case. A given text can be used in a multitude of different activities (fill in the blanks exercise, comprehension activity…) with as many different audiences, end user roles and typical learning ages. The same remark can be said about the element "Typical Learning Time" (which corresponds to the GEM element "duration"). It highly depends on the activity using the text and the audience to which it is submitted. There is no need to go through all data elements, most of them follow the same pattern, that is to say describe the object through its use. A text having many different potential uses, they do not constitute a good basis of description.

Additionally, we can argue on the usability of some of the elements value spaces. For instance the element "Interactivity Level" is not only unsuitable to our problem – for the different activities that can be supported by a given text can have various levels of interactivity – it is difficult to fill for lack of guidelines. If a teacher has to fill this element he (or she) will have to choose between "*very low, low, medium, high, very high*".

5. Leads towards pedagogical indexation

The existing standards in the field of pedagogical resource description are not adapted to our problem, which is too specific for them to be applicable. Their "*will to integrate within the same model entities of conceptually very different levels*"[9] is, as suggests Pernin for LOM, at the origin of their imprecision and ambiguities. Pedagogical indexation of texts for language teaching is too specific a problem to be handled directly and solely by LOM or any of the other standards presented. An ad hoc solution has to be implemented.

5.1. Metadata vs. document analysis

The imprecision in LOM mainly concern the data elements of the "Educational" category. Some of the other elements can be used as such in our system for they refer to inherent characteristics of texts, such as the author, the title, the year of publication etc… It is therefore not excluded to implement a metadata application profile: "*an assemblage of metadata elements selected from one or more metadata schemas and combined in a compound schema*"[10].

Nevertheless, metadata will not be sufficient to describe pedagogically a text. The fact that "pedagogical" data elements do not correspond to inherent characteristics of a text does not mean that the information they could have contained would not be useful for teachers. To be consistent, the description of a text description needs to contain exclusively information that is accurate whatever the use of the text might be. We therefore propose to complete the information provided by the metadata using a two phase process. The first phase will consist in annotating within the text characteristics that are relevant to the teachers but that will not depend on the exploitation of the document, such as lexical or syntactic characteristics (lemmas, forms…). These characteristics can to some extent be annotated automatically by NLP tools such as a lemmatizer or a morpho-syntactic analyser. Such information is mandatory to allow such queries as one for texts in English containing multiple occurrences of preterit. Teachers could need to precise the previous query by stating that most verbs need to be irregular or that they want some of them to be used in contrast with pluperfect occurrences. Unfortunately, annotation as we described it is probably not going to be sufficient to model all the queries of our group of users (cf. [3]). An additional software layer will address this problem. Ideally it would be an inference engine analyzing the different characteristics of each text (both metadata and annotations) combined with information provided by the user's query to deduce certain qualities of a text under the scope of how it is going to be used.

5.2. Facets

Considering the use cases of this system, the inference engine will just be left to handle what it can without adding errors (noise or silence), the object of the project is not to try to model the teacher's text choosing process, but to provide them with a text search aid. Such an aid will be provided by offering them the possibility to consider the different facets of a text, depending on their queries. The system aims at presenting them a subset of texts satisfying a query and eventually helping them to visualize certain aspects of these texts (such as the presence of certain grammatical structures, vocabulary…) thanks to the use of NLP tools. This presentation of the texts through their different facets is meant to grant the teacher access to aspects of the text he or she would have needed to read the whole document to be aware of. Additionally they will allow the global consideration of the text considering each facet quantitatively (are the sought elements numerous?) as well as qualitatively (which are they?).
The next phase of the project is to select which facets to make available based on their needs (cf. questionnaires) and the tools that can be made available to them (whether they are NLP tools or not).

References

[1] Antoniadis G., Echinard S., Kraif O., Lebarbé T., Loiseau M., Ponton C, CALL: from current problems to NLP solutions, in Proceedings of EUROCALL, Vienna, Austria, 1-4 September 2004
[2] Antoniadis G., Echinard S., Kraif O., Lebarbé T., Loiseau M., Ponton C, NLP-based scripting for CALL activities, in Proceedings of Coling '04 Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning, Geneva, Switzerland, August 2004
[3] Lefèvre P., La recherche d'informations, du texte intégral au thésaurus, Hermès, Paris (2004)
[4] Loiseau M., La description de ressources pédagogiques : état de l'art et application aux ressources textuelles pour l'enseignement des langues, Proceedings of workshop "TAL et apprentissage des langues", Grenoble, France, October 22 2004, http://www.u-grenoble3.fr/lidilem/talal/actes/JourneeTALAL-041022-loiseau.pdf
[5] DCMI Usage Board, DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms/ (2005)
[6] GEM Top-Level Elements, http://www.thegateway.org/about/documentation/metadataElements/ (2004)
[7] EdNA Metadata Standard V1.1., http://www.edna.edu.au/edna/go/pid/385 (2002)
[8] Final 1484.12.1 LOM Draft Standard Document, http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf
[9] Pernin, J.P., A propos des objets pédagogiques, in "Entre technique et pédagogie : la création de contenus multimédia pour l'enseignement et la formation", IRDP, Neuchâtel (2004), ISBN 2-88198-010-4 http://www-clips.imag.fr/arcade/User/jean-philippe.pernin/recherche/download/Article_Pernin_Neuchatel07Nov03.pdf
[10] Duval, E., Hodgins, W., Sutton, S., Weibel, S. L. Metadata Principles and Practicalities in D-Lib Magazine, **8 (4)**. (2002) http://www.dlib.org/dlib/april02/weibel/04weibel.html