MODEL FOR PEDAGOGICAL INDEXATION OF TEXTS FOR LANGUAGE TEACHING

Mathieu Loiseau, Georges Antoniadis and Claude Ponton
LIDILEM, Université Stendhal Grenoble 3, Grenoble, France

Keywords: Computer Assisted Language Learning (CALL), Natural Language Processing (NLP), Educational Metadata.

Abstract: In this communication we propose to expose the main pedagogical ressource description standards limitations for the description of raw ressources, through the scope of pedagogical indexation of texts for language teaching. To do so we will resort to the testimony of language teachers reagarding their practices. We will then propose a model supposed to exceed these limitations. This model is articulated around the notion of text *facet*, which we introduce here.

1 CORPORA FOR LANGUAGE TEACHING

Thanks to the communicative approach's widespread use (cf. (Levy, 1997)) authentic text is at the heart of the teachers set of problems.
However, corpora, despite numerous, are not dedicated to text search for language teaching. Querying mecanisms mostly rely on traditional keywords queries. Teachers display an ability to adapt computer tools of which they were not meant to be the end user, such as in Tim Johns' Data Driven Learning (DDL)(Higgins and Johns, 1984). All the same, some of the flaws of CALL systems mentioned in (Antoniadis et al., 2004) accurately describe the situation of language corpora for language teaching: when a teacher seeks to find a text in a corpus , there is no system that allows him/her to express his/her request in terms of his/her set of problems, using pedagogical concepts.

1.1 Pedagogical Indexation for Language Teaching

The project to create a pedagogically indexed text base for language teaching directly stems from these considerations. This project will lead to the implementation of a prototype (under development). It should fulfill the following use cases: text query and text addition.

Lefèvre's definition of "documentary language"[1](Lefèvre, 2000) explicitely puts the users at the center of the indexation process. Consequently:

Definition (Pedagogical Indexation). Pedagogical indexation *is performed following a documentary language, which describes objects according to pedagogical criteria (relevant to didactics).*

In our project, the considered objects are texts and we want the users (language teachers) to be able to find those objects by formulating questions that are relevant to their set of problems, *i.e.* language didactics. The scope of this article is that of pedagogical indexation of texts for language teaching.

2 USERS' NEEDS

In order to define pertinent criteria for text pedagogical indexation for language teaching, rather than favoring a given teaching approach, we have adopted an empirical approach: a preliminary qualitative study based on eight interviews with language teachers; a short questionnaire destined to grasp how teachers handle authentic texts, the classification and research of texts and to validate the apparently self-evident -

[1]"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*", translated by the authors

Loiseau M., Antoniadis G. and Ponton C. (2008).
MODEL FOR PEDAGOGICAL INDEXATION OF TEXTS FOR LANGUAGE TEACHING.
In *Proceedings of the Third International Conference on Software and Data Technologies - ISDM/ABF*, pages 212-217
DOI: 10.5220/0001893202120217

yet fraught with consequences - hypothesis that a variety of pedagogical contexts[2] can correspond to one given text; and a long questionnaire aiming at fine-tuning the information gathered in the first questionnaire and isolating research criteria.

2.1 Questionnaires Results

First Questionnaire. 112 out of 115[3] language teachers declare "being able to use a same text in various different contexts". It is not only prospective thinking since 106 of these 112 teachers declared having actually done it. Besides the confirmation of our hypothesis, we concluded that: teachers favor the use of authentic texts[4] ; they resort to specially constructed texts[5] when they want to control their linguistic content (grammatical structures, vocabulary), particularly for beginners groups; the type of activity and the audience seem to be the most frequent and common research criteria; finally, we could not to draw conclusions concerning the teachers' own text collections organization.

Second Questionnaire. We have been able, through the description of the teachers' own text collection classification, to isolate some research criteria, the most widely used of which were: theme, "linguistic content[6] or objective[7]" and level.
Aside from assessing the teachers' expectations towards a pedagogically indexed text base, the rest of the questionnaire was dedicated to confront the teachers' practices with the hypothesis that some criteria influenced one another. We have been able to demonstrate that the activity type had an effect on:

- text length, [F(5,143)=3,362; p<,01][8];
- the number of "representative elements" of the notion the activity is about, [F(4,127)=4,739; p<,005][8];
- the amount of unknown structures (other than the object of the class), [χ^2=32,177; dl=10; p<,001][9];
- the amount of unknown vocabulary (other than the object of the class), [χ^2=28,949; dl=10; p<,005][9].

Moreover, teachers declare that the level of the students influences the quantity of unknown structures tolerated in a text. They also state that the students' native tongue, when taken into account, has an influence on the quantity of unknown structures tolerated in a text.

2.2 Conclusions

A given text can be used in various different pedagogical contexts. For instance, the quantity of unknown structures and vocabulary sought (or tolerated) depends on the activity type. This "*un*-knowledge" is part of the learner's level. Therefore, we can say that, depending on the type of activity, a given text can correspond to more than one audience.
In our opinion, this example illustrates the fact that some "pedagogical" characteristics of a text are not fixed characteristics. They depend on the combination of some inherent characteristics of the text (such as its linguistical content) and on the pedagogical context in which the teacher plans to use it.
Our prototype will not cover all the teachers' needs and those which will be covered will only be partially so, hence the need for evolutivity.

The system will not be able to provide only the one most relevant text for every query. Whereas, it would not be a problem to fetch a text, based on its author and title; for other types of request, some criteria will be too hard to model.
Let us consider the example of the "theme" of the text. There is no reliable NLP tool that would extract the various themes of the text. Manual annotation would raise the issue of consistency and exhaustiveness. Even if the annotation in itself was consistent and exhaustive, the issue of linking the annotation with teachers queries remains. This type of criteria thus requires a certain number of approximations both for the indexation and for the expression/interpretation of the queries. Other criteria such as "how interesting the text will be to the students" are, to this moment, almost impossible to model. They influence the teachers' choices all the same. Our system cannot substitute for this choice process, it can facilitate it: act as decision-assistance by providing a subset of candidate texts highlighting certain elements of the text depending on the query that was performed.

[2] by "*pedagogical context*" we mean the didactical goals and all the characteristics of the audience (level, age, interests, etc.) and of the institution (track/diploma, material constraints, number of learners, etc.)

[3] The corresponding question was not on the first paper version of the questionnaire, which 18 persons answered

[4] in (Taylor, 1994) Taylor quotes various consistant definitions of "authentic text", among which Nunan's: "*A rule of thumb for authentic here is any material which has not been specifically produced for the purposes of language teaching.*"

[5] Unlike authentic texts, specially constructed text are written for the purpose of being used as teaching material.

[6] of the text

[7] of the activity the text is to be used in

[8] Anova test results, for information only

[9] χ^2 test results, for information only

3 "FLAWS" OF PEDAGOGICAL RESOURCE DESCRIPTION STANDARDS

Pedagogical resource description standards do not seem to bring an answer to the needs mentioned in the previous section. In this article, we will only explain our remarks concerning Learning Object Metadata (LOM) but they can be extended to most educational metadata standards.

3.1 LOM

LOM specifications propose more than seventy data elements. For each data element, the LOM specifications(IEEE, 2002) define the meaning of the data element, a cardinality, a datatype - sometimes accompanied by a vocabulary, which despite examples can be vague and bound to construct its meaning through the experience of the community of users. For instance, data element 5.3, "Interactivity Level" will take its value among *very low*, *low*, *medium*, *high*, *very high*.

Each element is non-mandatory and can be repeated to describe a "pedagogical object": "*any entity - digital or non-digital - that may be used for learning, education or training*".
LOM descriptor 1.8 ("Agregation Level") and its four levels, spanning from raw resource to fully integrated curriculum, illustrates this definition of pedagogical object. Pernin justly remarks that the ambiguities and imprecisions of the model stem from their "*will to integrate within the same model entities of conceptually very different levels*"(Pernin, 2004). (Translation by the authors)

3.2 Inherent Pedagogical Properties

All these metadata standards rely on the same principle: making an inventory of the properties of an object. The essence of this metadata introduces a notion of fixedness of the properties, including pedagogical properties. If fixed pedagogical properties can adequately describe objects that are already pedagogically exploited[10], it does not seem fit for raw resources, *a fortiori* texts.

LOM category 5 (Educational) descriptors designate either properties we have presented as pedagogical context, or properties, which are linked to it: 5.10, "Description", which contains "*Comments on how this learning object is to be used*"; 5.7, "Typical Age Range", the "*age of the typical intended user*"; and the self explanatory 5.8, "Difficulty". We have shown that a text could have various uses. Each different use (5.10 Description), will also affect both the object's difficulty and the audience to which it is destined. LOM allows various "5. Educational" groups. In our case an exhaustive description of the resource would require all possible uses of every text to be given in their LOM description. For each text, the "Educational" description would be an n-tuple, each component of which would be a set of "Educational" data elements.

Under these conditions, a manual pedagogical description of texts is too tedious to imagine thus indexing them, might it be with LOM or any other pedagogical resource description standard.

[10]By "pedagogically exploited", we mean that they benefitted from some pedagogical added value, which can be through the creation of an activity or other type of pedagogy oriented commentaries

4 THE NOTION OF "FACET"

Given the definition we provided, we have proved that the pedagogical indexation of texts for language teaching can be difficult to achieve if one considers all pedagogical properties as inherent to the object. To consider the various parameters influencing the properties of the text, we suggest the notion of "text *facet*" :

Definition (Facet). *A* text facet *is a property defined with a view to the latter's pedagogical exploitation in laguage teaching, accompanied by at least one mechanism to compute (automatically or not) the value of this property for any text depending on a given pedagogical context.*

The creation of a facet stems from the texts' subsequent pedagogical exploitation. Any characteristic that can be useful to a language teacher for the search or the choice of a text can be a facet of a text, provided that its value can be assigned or computed. The computation of a facet value might involve any text charasteristic. Those characteristics do not need to be pedagogically relevant, as such.

4.1 Facets and Parameters

The difference between facet and element, as defined in metadata standards, resides less in its semantics than in the means to obtain and use them. A parametrized facet could be one that numbers the representative elements of a language didactics relevant notion present in a text. From now on we will designate this facet by F_{RepE^t}. The value of this facet, is contextual and depends on the pedagogical context:

the notion on which to work. A teacher of Spanish might find it interesting to work on the expression of duty, for instance with haber* que + Inf [11] or tener* que + Inf.[12]. Obviously a text will not necessarily have the same number of occurrences of both structures, the value of F_{RepE^t} will thus be different according to which structure (pedagogical context in this case) is sought. We will discuss in 5.2 how to make these facets available to the user.
To implement F_{RepE^t}, a morphological parser (NLP) can be combined with a pattern matching program (to find the sought structure) and a counter.

Constant Facet. In the same way the constant function $f(x) \rightarrow 0$ is a function of x nontheless, a facet assigning the same value to a text for any pedagogical context is all the same a facet.The length of the text (F_{length}) or its author are thus facets. These facets, like pedagogical resource description standards metadata, consider properties that are intrinsic to the text.

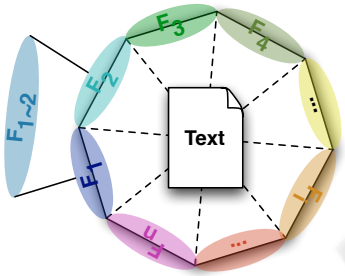

Figure 1: Text facets.

Compound Facets. All the previous examples either rely on user submitted information (cf. Author) or on underlying text properties revealed by computer programs (including NLP programs). It is also possible to combine various existing facets to create others with higher pedagogical added value (Figure 1).
Let us combine two facets we have already discussed: F_{length} and F_{RepE^t}. We can perform an approximation of a facet evaluating the adequation of a text for a pedagogical context depending on two parameters: the activity type and the notion at the center of the activity (2.1). Since we have shown that the activity type both had an effect on text lengths and the number of representative elements, we can reuse the data from our questionnaire to establish threshold values depending on the activity type.
Figure 1 emphasizes the fact, that combining two facets is different than just giving both a value. In this case the use of the compound facet infers from the activity type the desired number of words of the text as well as the number of representative elements of the notion (along with the tolerance). Of course, this data comes from the teachers' *declared* practices, it is therefore likely that the facet will not be very powerful. Though, studying the *actual* practices with the prototype should allow a refinement of those values and the iterative creation of more powerful facets. For instance, this particular compound facet could then be improved, by taking into account the students level.

[11] **Hay** que comer / **One** must eat. (Impersonal)
[12] **Tienes** que comer / **you** must eat.

4.2 Facet *vs* Descriptor

These examples evoke the major difference between metadata descriptors and facets. The descriptors contain a set of *a priori* defined fixed values. Whereas the facets' can also be dynamically computed according to a pedagogical context entered by the user.
For some simple facets, one could imagine a mechanism that would create the Educational elements of an application standard of LOM. For a "level" facet that would take into account the activity type, it would be possible to create an "Educational" tuple for each possible pedagogical context. But other facets, such as the F_{RepE^t}, depending on the formalism used to describe the notions, the number of possible pedagogical contexts (hence the number of tuples) would make exhaustiveness an issue.

5 THE MODEL

Facets allow to integrate processes to the description of resources, without actually containing them.

5.1 Prism

The revelation of a facet requires certain processes which can be performed by a human (*e.g.* "author" facet) or by a program. Hence the notion of prism:

Definition (Prism). *The* prism *is a set of computer function(s) created or combined in order to reveal the facets of a text.*

The functions in Figure 2 ($Fn_1 \rightarrow Fn_n$) are inspired from the MIRTO model(Antoniadis et al., 2005): in both cases the functions in themselves do not have a didactic interest. Their combination at the level of the prism (respectively scripts) is what creates the pedagogical added value. Tokenization is not a functionality of the prism, but the succession of tokenizer and counter is: it gives access to F_{length}.

In Figure 2, one can see two facets to each of which a treatment in two parts is associated. The first

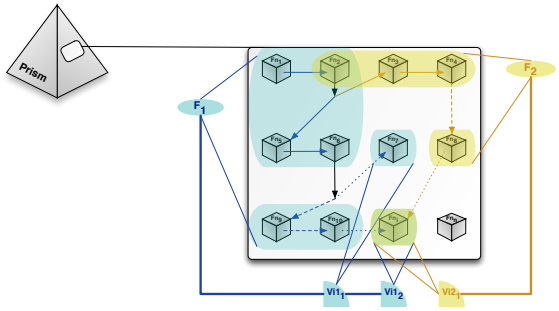

Figure 2: The prism.

part of the treatment is to be performed once and for all when the text is added to the system. It is meant to add the underyling data from which the values of the facets are computed. The second part is the actual process which computes the values of the facets using the pedagogical context. The processes associated to two different facets can both use the same function (Fn_2, in Figure 2, *e.g.* a counter).
The creation and the modification of the prism require computer science expertise which cannot be expected from a language teacher. This task ought to be performed by developpers, but cannot be so without language didactics knowledge considering the pivotal position of the prism in our model. We thus advocate the inter-disciplinary dialogue with teachers and language didactics experts for this phase.

5.2 Access to the Facets of a Text

To make the prism revealed facets available to teacher users, we propose two non exclusive means: views and visualizations.

5.2.1 View and Empty View

Facets are text properties, revealed thanks to the conjunction of computer programs combined in the system prism. The functions of the prism not only allow the necessary processes for revealing facets, they are also used to perform queries on these values.

Definition (View). *The* view *of a text t according to the facet F under the pedagogical context C, is the value of F computed for t using the characteristics of C.*

In other words, given a facet and a text, the prism computes a different view of this text for each pedagogical context specified by the teacher through the system interface.

By extension, the view of a text set is a set of the respective views of these texts. By constraining the values of the facets, thus extending the pedagogical context to the expected values, the view of a set of texts acts as a filter: the view of a text is "*empty*" if it does not satisfy the constraints.
In F_{RepE^t}, there is a view for each language fact of which we might want to count the occurences. Let us use again the pedagogical context haber* que + Inf. A view of a text for the facet and pedagogical context is the number of occurences of haber* que + Inf. A constrained view of the same text, using the same facet and extending the pedagogical context, could impose that the corresponding view is greater than a given number: *e.g.* the view of F_{RepE^t} for the context count of "haber* que + Inf" structures ≥ 4. The view of a text, that does not satisfy this constraint will be considered an *empty view*. By extension the view of a text set will contain all the non-empty views of the texts of the set. *I.e.* the views for all the texts having at least 4 occurences of haber* que + Inf structures in this case.

5.2.2 Homogeneous Text Collection

The notion of "homogeneous text collection" directly stems from that of view of a group of text.
A text collection, will be considered homogeneous if all the texts it contains explicitly share one or more properties.

Definition (Homogeneous Text Collection). *An* homogeneous text collection *is a constrained view of a text collection.*

The homogeneity of the collection therefore depends on the satisfaction of the research criteria. This notion allows to perform a research on a subset of texts: the user will be able to refine his research step by step. Each step being the creation of a view of the previous step (Figure 3).

These intermediary collections will allow better performances than adding parameters to the query and performing it on the whole collection all over again.

5.2.3 Visualizations

To provide the user with text choice assistance (cf. Figure 3), the system provides the user with graphical representations of the underlying information added by the prism in relation with a given facet. These representations are called "*visualizations*". The information of the visualizations is qualitative, whereas the view is quantitative. On Figure 2, F_1 is associated to two visualizations: Vi_1, uses the underlying information associated to the facet and Vi_2 uses the view itself.

To help the user choose between the texts satisfying a new view of F_{RepE^t} - texts with 5 occurences of English Preterite - the system could provide a visualization highlighting each occurence or another one,

which would only show the list of the occurences of the structure (Figure 4).

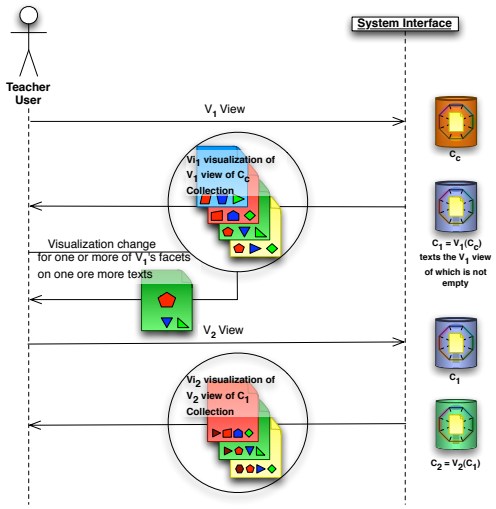

Figure 3: Interaction diagram.

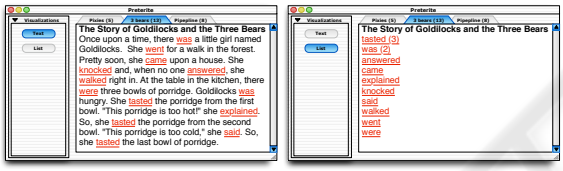

Figure 4: Visualization examples.

6 CONCLUSIONS

A study of language teachers' needs and practices proved pedagogical resource description standards inadequate to describe raw resources, in particular texts, in the context of pedagogical indexation for language teaching. We have presented a model, which attemps to take into account the influence of the pedagogical context on the value of the texts' properties. To do so we introduced the notions of facet, prism, view, visualization and homogeneous text collection.

The implementation of such a system must be as generic as can be: its quality will strongly depend on its ability to integrate different kinds of tools and on the quality of the latter. While writing these lines, we are implementing the architecture we have described using php/mySQL. The prism of the prototype will integrate a few simple tools to grant access to a few basic facets. The prototype is meant to be an implementation example of the architecture, showing its viability. But any urge to make a product out of it would require several distinct tasks. The analysis of the teachers needs: the study of the teachers' *actual* practices would refine our declared practices based asumptions of the teachers' needs; such a study could use our prototype. Isolating text properties from a language didactics point of view that would prove useful to the teachers. And finally, developing reliable NLP tools, which would add information to the texts that can be used for facet creation. Of course, each these three leads must feed off the conclusions of the other two.

REFERENCES

Antoniadis, G., Échinard, S., Kraif, O., Lebarbé, T., Loiseau, M., and Ponton, C. (2004). Nlp-based scripting for call activities. In *Coling Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning Proceedings*, Genève.

Antoniadis, G., Échinard, S., Kraif, O., Lebarbé, T., and Ponton, C. (2005). Modélisation de l'intégration de ressources tal pour l'apprentissage des langues : la plateforme mirto. *ALSIC*, 8:65–79.

Higgins, J. and Johns, T. (1984). *Computers in language learning*. Collins ELT / Addison-Wesley, World Language Division, London.

IEEE (2002). Final 1484.12.1 lom draft standard document.

Lefèvre, P. (2000). *La recherche d'informations : du texte intégral au thésaurus*. Hermès Science, Paris.

Levy, M. (1997). *Computer-Assisted Language Learning, context and conceptualization*. Oxford University Press.

Pernin, J.-P. (2004). À propos des objets pédagogiques. In *Entre technique et pédagogie : la création de contenus multimédia pour l'enseignement et la formation*, Neuchâtel.

Taylor, D. (1994). Inauthentic authenticity or authentic inauthenticity? *Teaching English as a Second or Foreign Language*, 1(2):A–1.