“Facets” and “Prisms” as a Means to Achieve Pedagogical Indexation of Texts for Language Learning: Consequences of the Notion of Pedagogical Context

Mathieu Loiseau, Georges Antoniadis, Claude Ponton

"Facets" and "Prisms" as a Means to Achieve Pedagogical Indexation of Texts for Language Learning

Consequences of the Notion of Pedagogical Context

Mathieu Loiseau, Georges Antoniadis, and Claude Ponton

LIDILEM, Université Stendhal Grenoble 3
BP 25, 38040 Grenoble cedex 9, France
{mathieu.loiseau,georges.antoniadis,claude.ponton}@u-grenoble3.fr
http://w3.u-grenoble3.fr/lidilem/labo

Abstract. Defining pedagogical indexation of texts for language learning as an indexation allowing users to query for texts in order to use them in language teaching requires to take into account the influence of the properties of the teaching situation we define as "pedagogical context". We propose to justify the notions of prisms and facets on which our model rely through the description of material selection in the task of planing a language class as an adaptation of Yinger's model of planing. This interpretation of Yinger's model is closely intertwined with the elaboration of the notion of pedagogical context. The latter provides sounder bases on which to build our model on. This resulted in improvements in the potentialities of the model compared to its first published version.

Keywords: Pedagogical indexation, Computer Assisted Language Learning, Natural Language Processing, Metadata, End User Programming

1 Pedagogical Indexation

The MIRTO project, started in 2001, stemmed from the observation of various recurrent issues in Computer Assisted Language Learning (CALL) systems: rigidity, inability to adapt the learning sequences to learners and unavailability of means to manipulate concepts pertaining to the teachers' field of expertise (language didactics) [1]. The aim of MIRTO was to promote the use of Natural Language Processing (NLP) to address those problems by adding an abstraction layer between the user and the material. Antoniadis *et al.* consider that the formulation of problems in didactics relevant terms depends on handling language not as character sequences but as a system of forms and concepts [1]. MIRTO thus proposes to separate treatments (*e.g.* gap-filling exercise generation script) and the data on which they are to be applied (a text in this case).

1.1 Definition and Objectives

This made evident the need for a text base, which, for consistency's sake, would have to allow users to perform language teaching driven queries. In other words,

a subpart of the problem was the conception of a system that could perform pedagogical indexation of texts. In this work we defined pedagogical indexation as "an indexation performed according to a documentary language that allows users to query for objects in order to use them for teaching" [10, p. 15]. Considering the aforementioned context, we are therefore working towards pedagogical indexation of texts for language learning.

Indeed, a study of the literature concerning the most often used language teaching methods and a series of interviews with some language teachers prompted us not only to consider this problem in the context of the future use of the text in a CALL activity, but to try to consider the problem globally: few of the teachers we had interviewed were really computer savvy, all the same, they all underlined the importance of text search in their practices. We later got confirmation of this nature of things by a larger scale study, which established text search as a common task in language teaching [10, p. 170].

Having modified the scope of our work – without completely cutting ties with MIRTO, for integration remained a perspective – into the conception of a model for pedagogical indexation of texts for language teaching, we started to consider the existing means to achieve it.

1.2 Learning Resource Description Standards

A wide array of research tackles the definition and use of learning resource description standards. The principal standards we analyzed were LOM [9], SCORM [17] and some teaching oriented application profiles of the Dublin Core (edna[4] and GEM[5]). As for providing a solution to our problem, all the standards we studied came with the same flaws, most of which stem from the fact that these standards try to integrate in the same model, entities of very different conceptual level: the ressources used to set up activities (low aggregation level in the LOM terminology) and the activities themselves (higher aggregation level) [14, p. 2]. Balatsoukas *et al.*take this analysis further in pointing out that the lower the aggregation level of the learning object the broader its spectrum (*i.e.* the range of activities that can be performed with it) [2]. Indeed, in the particular case of texts (raw resources), the descriptors provided by the standards seem, at best, difficult to use: how does one assign a "Description" ("Comments on how this learning object is to be used" [9, element 5.10]) when the resource potentially could be used in different contexts.

The approach advocated by Recker & Wiley proposes to treat differently what they call intrinsic ("derivable by simply having the resource at hand") and extrinsic properties (which "describe the context in which the resource is used") [16, p. 260]. All the same, their analysis cannot be directly transposed to our problem, for their aim is to provide a collaborative resource description system in which authoritative and non-authoritative annotation coexist. On the other hand our aim is, in the first place, to provide a model that would allow a system to automate as much as possible the pedagogical indexation of texts. User annotation is, in this context, more a potential extension of the system than a core feature. There was therefore at this point no clear cut direction in which to

go: the pedagogical properties seemed to constitute extrinsic properties for the raw resources that are texts, thus potentially discarding educational metadata as a solution. We therefore decided to resort to an empirical study to confirm this hypothesis and get a grasp of teachers practices regarding text search.

2 Pedagogical Context

Our empirical study took the form of a survey, which built on a series of interview and an exploration of the literature, part of which we have just summed up above. Beyond the confirmation of the hypothesis of the multiple uses texts can have in language teaching, we aimed at obtaining a first look into the process of text search. We meant our point of view to be as general as can be, in the hope to extract invariants that would remain unaffected by variables such as the language taught, the country in which it is taught or to whom. The study was mostly filled online, but also in paper form, both medium adding up to 130 testimonies. Beside confirming unequivocally that texts can be used in various language teaching situations[1], the survey allowed us to extract a (non necessarily exhaustive) list of four practices that lead to texts being used in language learning: search for a text to use in a precise activity, writing the text, text encounter during personal readings and texts from a syllabus (of any form). We will focus here on the provenance that is closest to the role of a pedagogically indexed text base, *i.e.* the search for a text in order to use it in a specific activity, which also happens to be the most widely represented practice (concerning nearly 97% of the teachers answering the survey).

2.1 Adaptation of Yinger's Model

To describe the task of searching for a text for a given activity we resorted to using Yinger's model of planification [18] or more precisely part of it. Yinger defines planing as a three stage process: problem finding, problem formulation/solution and finally implementation, evaluation, routinization [18, p. 30]. In our task, the problem is already found (the teacher has an activity in mind) and the search is supposed to provide a text to actually use in class and thus precedes implementation. We focus here on the problem formulation/solution, which according to Yinger is an "helicoidal" repetition of three phases: elaboration, investigation and adaptation [18, p. 35], which we adapt to our problem under the labels selection, evaluation and transformation (cf. figure 1) [10, pp. 205–210].

The dashed semi-ovoid at the bottom of figure 1 contains a set of texts the teacher has access to. The intensity of the gray inside the form represent to which extent they are pedagogically "connoted". For instance a text taken straight from a newspaper and that has never been used in teaching (to the knowledge of the teacher) is not connoted, whereas a text recommended by

[1] 97,3% of the teachers who answered the question declare they consider that a given text can be used with various goals in different contexts and 94,5% of them (92% of our the sample) declare having done so.

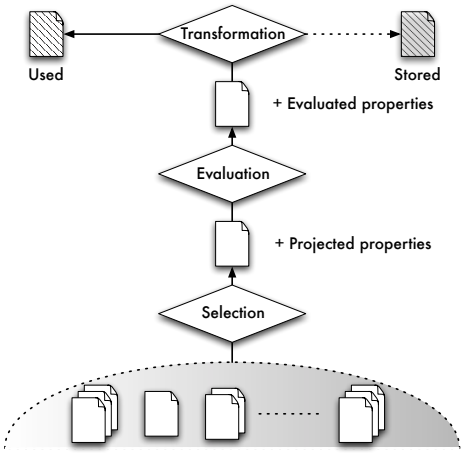

Fig. 1. Yinger's model adapted to text search

peers or found inside a textbook has some sort of pedagogical connotation. The aim is not to evaluate this "connotation" or even theorize it, but to acknowledge that the teacher can resort to sources with different statuses.

The selection phase consists in the teacher relying on his necessary preconceptions[2] projecting onto the text properties linked in a way or another to the activity they are planning. An example of such a behavior is choosing an author based on properties attributed to their writing: "Roald Dahl, [...] all his short stories are packed with these verbs [...] for emotion and gestures [...], that in French [require] a whole phrase [...]." (testimony from our study).

Once the text is selected, based on the properties that the teacher has attributed *a priori* to it, it is actually in the hands of the teacher (or virtually so) for the first time in this planning sequence. They can now attribute a new set of properties to the text. The latter are no longer projected properties, they constitute the teacher's actual perspective on the resource based on the activity they want to set up with it. This set of properties can confirm or invalidate the ones that have been assigned during the first phase or concern totally different aspect of the text. For instance, it is completely imaginable that the teacher we quoted above should confirm her hypothesis, but conclude that the short story can turn out to be difficult for her learners, which brings us to the last phase: taking action upon the evaluated properties. The action transforms the text status-wise, there are three alternatives:

- the text is assigned a use context corresponding the teacher's current search and is transformed into actual teaching material (solid arrow in figure 1);
- the text, though considered unfit for this particular activity, is deemed useable in another context and can be kept for future use in a personal repository: it is transformed into potential teaching material (dotted arrow);

[2] Without preconceptions this phase would consist in a random selection of texts.

– the text is not relevant from the teacher's point of view and is just discarded (not represented).

2.2 First Definition

The description of these three phases allowed us to precise the role of a pedagogically indexed text base: it is meant to assist the teacher in the selection phase and possibly allow him to perform it according to less instinctive criteria when applicable (for example concerning the linguistic content of the text), but it also allowed us to introduce the notion of Pedagogical Context (PC)[3] as: "set of features which describe the teaching situation" [12, p. 487]. This notion is especially useful in order to describe the process of text search and its integration in a learning sequence for the various iterations of the above scenario correspond to a gradual definition of the PC. The material is a component of the teaching situation [8, p. 31] thus influencing it. At the same time its choice is influenced by the other components of the PC since the search is performed for a given activity. In order to achieve pedagogical indexation of texts for language learning, it seems necessary to be able to take into account the PC, which means studying the link between components of the PC and the actual properties of the text.

3 Definition Update: Pedagogical Context as an Influence

Among our objectives with our second survey was trying to establish relations between properties of the PC and properties of the text. We cross-examined:

– the activity type (gap-filling exercise – 3 types –, comprehension activity, introduction of new notions – vocabulary or syntax –) with the size of text, the number of representative elements of a notion (if the notion is the preterit this will be the number of preterit conjugated verbs present inside the text, and the tolerance to newness (vocabulary and grammar-wise);
– the learners' first language and tolerance to newness ;
– the learners' level and tolerance to newness.

The length of the text and the number of representative elements were numerical variables and were asked for each activity type. In this case, the tolerance to newness was evaluated using two separate categorical variables, one concerning new vocabulary (other than the object of the lesson) and the other concerning new grammatical structures (other than the object of the lesson). Both variables could take their values between "proscribed", "tolerated" and "sought". For each activity type used, we asked the teachers to rate their tolerance to newness using this scale for both variables.

When crossed with the learners' level and first language, the tolerance to newness was also the object of a closed-ended question. These questions allow

[3] In order to avoid exceedingly numerous repetitions, we will either refer to it using "PC" or its complete form "Pedagogical Context".

the teacher to state that the criteria is not relevant or can decide not to answer. The other two possibilities depended on the question and do not distinguish vocabulary and grammar:

- first language : the more similar the mother tongue and the learned language, [*the more*/*the less*] one will accept unknown grammatical structures or vocabulary;
- level : the higher the level, [*the more*/*the less*] one will accept unknown grammatical structures or vocabulary.

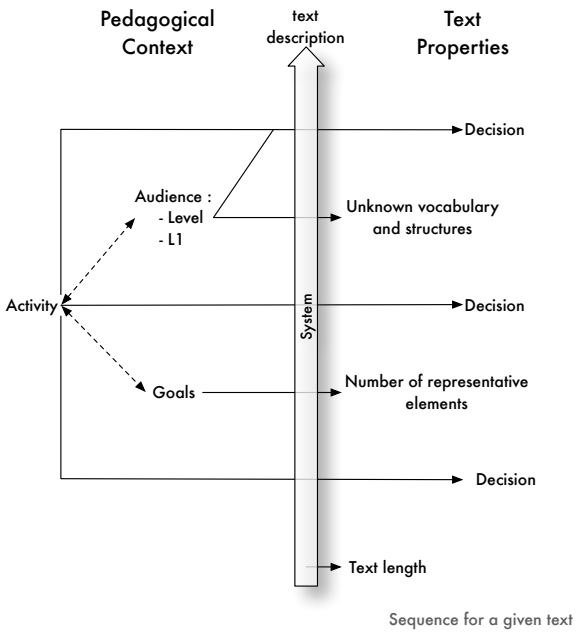

Fig. 2. Influence of the pedagogical context on the attribution of text properties

The results have been summed up by figure 2[4]. Properties such as the length of a given text are totally independent from the Pedagogical Context and thus do not need it to be computed, but our study showed that the activity type had an effect on text length[5], which means that depending on the activity type, teachers will be looking for texts of different lengths. A text property such as the number of representative elements of a notion obviously depends on the notion, which in turn is a direct consequence of the pedagogical goals of the teachers. Likewise, the number of representative elements of a notion considered appropriate by

[4] Due to room restrictions we cannot include detailed statistics in this paper, they are available in section 5.3 (pp. 231–245) of [10] though.

[5] ANOVA: $F(143) = 3,362$; $p <, 01$. Post-hoc tests are significant when comparing "comprehension activity" with the various forms of "gap-filling exercises" [10, pp. 238–239].

the teacher will depend on the activity type (*e.g.* 4 or 5 occurrences might be enough to introduce a notion, whereas to practice it under the form of a gap-filling exercise teachers seek an average of 11 occurrences)[6]. Finally, if the amount of unknown vocabulary/structures is a property of the text, it cannot be evaluated unless we link it with the audience with whom the activity is going to be used. It directly depends on the level of the students, which is also used differently afterwards to take a decision on whether or not to use the text: the higher the students' level the more tolerant the teachers will be regarding the presence of new vocabulary or structures (other than the object of the lesson). The activity type[7] and the proximity between the learners' language and the one that is taught also seem to have a significant effect on the tolerance to "newness"[8].

The various tests we have performed on the above series of variables tend to show that the Pedagogical Context indeed influences text properties. We lack data to precisely characterize the relations between text properties and the PC, but we have been able to demonstrate their existence. The fragmentary knowledge we have come to gather has allowed us to explore examples of ways to take into account these concurrent influences that the Pedagogical Context has on text properties or on the way to act upon them. Interestingly, they all follow the same pattern, the properties which depend on the PC represent a sort of point of view of the text reflecting the problem of the teacher in his search. The Pedagogical Context, despite still representing the same entities in the real world, thus becomes, thanks to this switch of focus, "a paradigm casting its influence on the texts' properties".

4 Prism-Facet Based Model

The following model aims at taking into account the role of the Pedagogical Context in the evaluation of text properties, in order to propose help to the user in his selection task. It is a second version of the model which has been introduced in [11]. We will first describe this new version of the model, before we conclude by explaining the main differences between the two versions.

4.1 Recursive Definitions

The model is articulated around a couple of two indissociable notions: prism and facet. The prism insures that the properties are coherent in the way they

[6] ANOVA: $F(127) = 4,739 \; ; p <, 005$. Post-hoc tests are significant when comparing "introduction of a new notion" with "comprehension gap-filling exercise" and "introduction of a new syntactic notion" with "form aimed gap-filling exercises" [10, pp. 239–240].

[7] $\chi^2(10) = 32,2 \; ; p <, 001$ [10, pp. 240–243].

[8] 81.3% of teachers taking into account their learners first language considered that closer languages allow more tolerance [10, pp. 243–244] and 71.4% consider that the higher the level of the learners the more unknown vocabulary/structures they will accept [10, pp. 244].

are computed: "a *prism* is a mechanism – computerizable or not – associated to a property defined considering the texts' later exploitation in teaching, which allows to assign a value to this property for all text depending on a given pedagogical context"[9].

This definition allows us to highlight the link with pedagogical indexation: the definition of the prism depends on the needs of the teachers. This definition revolves around the difference between the conceptual level of the properties (class of properties) and their value (after instantiation). It is the essence of the prism which is the procedure which allows to make the transition from the first to the latter, when applying the concept to a given object (a text).

This leads us to the formalization of the property. Like the prism depends on the property it is meant to describe, the latter depends on its *alter ego*: "a text *facet* is a property of the text, which was defined with a view to its pedagogical exploitation in laguage teaching and for which an evaluation procedure can be defined and applied to any given couple $(text, \mathrm{PC})$"[9].

4.2 Facet and Facet-Value

Before we go on and explore the consequences of the above definitions, we shall enter a terminological issue. Like the term "property", the word "facet" is, as we use it, polysemic. It can, depending on the context, designate either the concept or the attribute. For instance, "parallelism" is a property (concept) which is applicable to a certain type of object, and two planes (for instance the ground and a shelf) can have the property to be parallel. In the case of facets, we might use the word to designate either the property in its conceptual form – facet F_i, text facet or just facet with no other precision – or its value for a given couple $(text, \mathrm{PC})$ – a given text's facet, $F_{i[CP]}(T)$ –.

4.3 Constant Facets

From the point of view of the task of selection, the facet is the central entity on the conceptual level: in the planning process, the facets represent the notions upon which the teachers base their reasoning. A pedagogically indexed text base will not be able to take into account every teacher's individual point of view of every facet presented to them (or at least in the near future), the usability of such a system therefore relies on the prisms, which offer consistency through their mechanical, systematic, nature.

Going through some of the properties represented in figure 2 will allow us to explain further the model.

In figure 3, we indicate two examples of facets. The word count ($F_{wrdCount}$), which is exactly the same as the property in figure 2 and F_{author} corresponding to the author of the text. The diagram also presents the values of these facet for a given text T. We introduce a functional notation based on the facets, even though strictly speaking the *application* that allows the computation of the

[9] Translation of the definitions page 257 of [10].

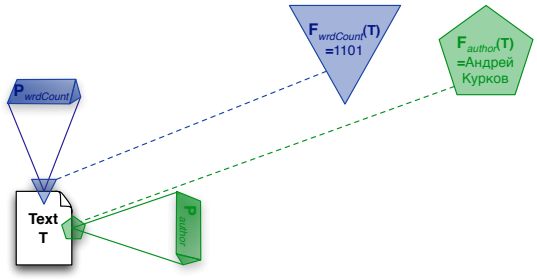

Fig. 3. Prism examples and values for the corresponding facets (PC independent)

values is defined inside the prisms ($P_{wrdCount}$ and P_{author}, here), which precise the status of both entities:

- the prism is a tool, materializing a process;
- the facet is a concept, a text property which has a value for every couple $(text, \text{PC})$.

4.4 The Pedagogical Context in the Model

In these first examples, the Pedagogical Context does not influence the value of the facet, which remains constant for a given text T for any PC. The aim of the model is to represent more complex properties. In figure 2, the number of representative elements is an example of such a facet. We represent it in figure 4. In this figure, a sole prism (P_{repE^t}) is shown revealing two facets for each

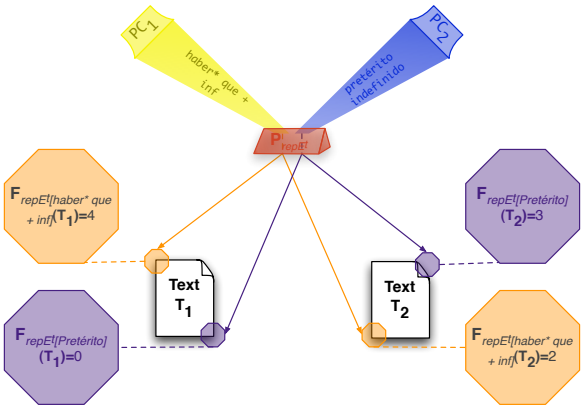

Fig. 4. Prism examples and corresponding facets for 4 different $(text, \text{PC})$ couples

text. Each of the two facets of T_1 and T_2 corresponds to a different Pedagogical Context for which both text could be compared, in order to come to a decision. In the example of figure 4, text T_1 contains 4 occurrence of *haber que* structures[10]

[10] Expression duty in Spanish: *Para soñar hay que dormir* (to dream, one has to sleep), *Habrá que resistir un tiempo más* (One will have to go on resisting for a while).

and no preterit, while T_2 contains 2 occurrences of *haber que* structures and 3 occurrences of preterit. figure 4 also represents the metaphor behind the name of prism and facets. In this metaphor the Pedagogical Context is a light cast on a text through a prism, thus revealing one of its facets. Consistently with its optical counterpart the prism divides the ray of the PC to keep only the components (frequencies) which are necessary to compute the value of the facet. Applied to a system which would assist the user in its selection task, the choice of prisms would have an expressive function: the user would only be asked to provide the PC components required by the prisms selected, thus providing them with means to describe the features of the teaching situation which are relevant for their search.

The notation introduced in figure 4 is meant to render the status difference that exists in the model between PC and texts. It comes from the function of the model, namely to provide a framework for the implementation of a system of pedagogical indexation of texts for language learning. When performing a given iteration of the cycle described in 1, the PC is constant. Of course, for a task of text search to yield a text that is actually used in language learning, the Pedagogical Context might evolve during the various iterations of the cycle, but the PC will be constant inside a given selection subtask (for which a system is supposed to provide assistance) of a given cycle. Yet, each prism is evidently meant to be reusable from one cycle to the other and, by definition, has to be able to compute values of its associated facet for all PC[11], hence the notation.

4.5 Prisms as a Means of Selection

By definition, indexation is essentially a description task [3, p. 419], yet it is aimed at allowing users to easily spot the texts that satisfy their needs, an objective of discrimination. In our case, part of the discrimination task, will not be automatable (e.g. based on *interestingness* or on the ability to give rise to a debate), the other part will mostly rely on constraining the tolerated values of facets. We have concluded that the better way to model that kind of constraint is to integrate it inside the Pedagogical Context and thus to take it into account in the value of the facets. A constrained version of a facet just adds a phase to the mechanism associated to its computation: after the value of the non-constrained avatar of the facet is computed a simple test instruction can be added, to return *false* if the constraint is not met and the value computed otherwise. In the constrained facet obtained, the expression of the constraint is part of the Pedagogical Context. Indeed, it is relevant to the problem of the

[11] The implementation of certain facets, such as the number of occurrences of a given type of reported speech (direct, indirect, free indirect) would require manual intervention. All the same a mechanism can be defined in order for a human to annotate it (making it a facet). In a system, such a facet could be implemented on a set of texts. To make such texts coexist in a system with not annotated texts (treating it as a subcorpus), *not applicable* has to be an accepted value of a facet for a text in a certain pedagogical context.

teacher to decide, depending on the situation they want to use the text in, to exclude texts based on the value of facets such as its length.

We have been convinced of that when trying to consider higher level facets. For instance, one can imagine developing a prism which would allow to take into account the information we have gathered in our study regarding the activity type[12]: let F_{AN} be the facet associated to this prism. F_{AN} could be a boolean property telling whether a text is potentially suitable for an activity. The PC components used would be the activity type and the notion on which to work. The treatment would rely on the facets we have called $F_{wrdCount}$ and F_{repE^t}, fixing threshold values for each activity type (for instance a gap-filling exercise could not be longer than n words and could not contain less than, say, 5 occurrences of the notion). The constraint of $F_{wrdCount}$ and F_{repE^t} is directly derivable from the PC of F_{AN}, which is a clue in the direction of our solution. But the decisive element is the fact that the threshold values that could be defined based on our study, despite lacking precision, come from teacher declarations. They were given the possibility to consider the criteria not pertinent, which means that it is very likely that it corresponds to a conscious feature expected in the text (if not explicitly evaluated) and thus qualify as a component of the Pedagogical Context. We do not consider this the only solution for devising a prism associated to F_{AN}, but find it a consistent and practical one.

4.6 Facet *vs* Metadata

The notions of facet and prism allow to:

- associate the concept (facet) and its modeling, making explicit the sense of the concept handled by the tool (prism);
- model the influence of the Pedagogical Context on the properties of the objects (texts).

These two characteristics distinguish facets from metadata. According to Bourda, metadata is information on objects which can be understood by humans and processed by software [7, pp. 116–117]. Both facets and metadata are therefore meant to propose a global point of view of an object rather than highlight information contained in the document (for instance F_{repE^t} means to provide a unique value associated to a structure, not to list all the occurrences of the structure). This similarity in the object of both notions is especially conspicuous for constant facets (cf. figure 3), which could be treated with metadata. But in the same way that constant functionals such as $f(x) \to 0$ are a particular case of functionals, constant facets are only a particular case of a generic notion, which cannot be efficiently modeled with metadata.

This can be shown with the example of F_{repE^t}. In order to implement comparable description with metadata one would need to anticipate any possible request made by teachers. The text "Rabbits run." would require a descriptor

[12] The actual implementation of such a facet would require much more experimentation: we only have *declared* practices, which would lack precision.

saying it contains one occurrence of the form "rabbits" but also one occurrence of a form the lemma of which is "rabbit". The text should also be found if the teacher is looking for the form "run", but also if they are looking for a text containing occurrence of the present simple of the verb to run. We already have 4 descriptors indicating one occurrence of a given structure. But it might also be pertinent to know that the text contains one occurrence of "rabbits run", one of a form whose lemma is "rabbit" with the verb run, one occurrence of the form run associated to a plural subject, etc. And this only concerns a 2 word text.

When the Pedagogical Context offers a certain variety of potential values – each of which should be associated with a value for each text – the fixedness of metadata requires to anticipate every single one of them, making it potentially hazardous or inefficient as far as storage is concerned (in our example, despite not being exhaustive, we have found 7 descriptors for a single facet and a two word text). Facets and prisms, by associating a property and a means to compute it introduce flexibility and dynamicity in the description of resources, which seem necessary to handle the notion of Pedagogical Context.

5 Towards Implementation

The example of F_{repE^t} leads to considering implementation options. Indeed, in order to introduce flexibility and to make computation of facet values possible, the information *on* the text provided by F_{repE^t} relies on information *of* the text. The computation of values of F_{repE^t} could be handled first by performing morphological analysis of the text, before using regular expressions on the resulting annotated version of the text. We will refer to the information of the text added by the first part of the process as underlying properties of the text. They are to be analyzed to provided information on the text, namely facet values.

When implementing this sequence of treatments in the perspective of indexing them, the addition of underlying properties (morphological analysis for F_{repE^t}), which will be referred to as pre-processing, should be performed once and for all, when the text is added to the system. On the other hand, in order to introduce the dynamicity that metadata lacks, the computation of facet values, which we will refer to as "computation", needs to be performed when the user queries the system.

5.1 Prisms and functions

This decomposition of the prism's mechanism as a sequence of treatments grouped into pre-processing and computation allows us to answer the question asked by note [11]: when implementing a facet based system, a prism mechanism can require human pre-processing but computation needs to be fully automatable.

As far as implementing prisms, to provide evolutivity and take advantage of already developed tools (especially NLP procedures), we recommend reusing the concept of function as defined in MIRTO [1]. According to this point of view a prism is linked to a facet and composed of two sequences of functions: pre-processing and computation (cf. figure 5).

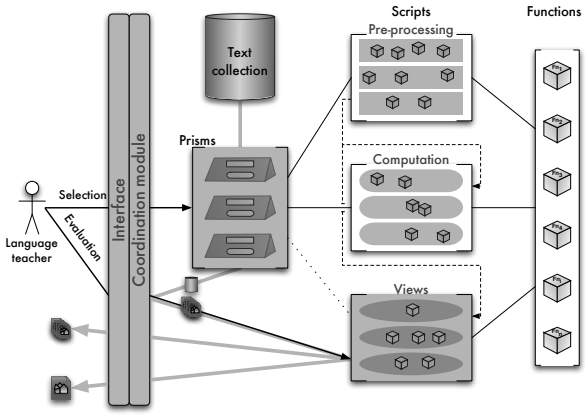

Fig. 5. Proposed general architecture for a facet based system

5.2 Views

In figure 5 prisms are not the only entity composed of functions. As an extension of the indexation system and a means for the user to interact with the system we introduce the notion of views. Considering the complexity of certain properties which intervene in the process of searching for a text to use in language teaching and the difficulty to achieve reliability in NLP when moving away from the form, a realist approach needs to acknowledge the amount of work left to the user during the phase of evaluation. Among other considerations, the fact that "100% reliability is, and may stay in the future, an unattainable goal. Therefore it is more realistic to stress on 'assisted' rather than 'fully automated' approaches" [6] is at the origin of Blanchard *et al.*'s "didactic triangulation strategy". Adapting it to our problem, views come as a means to assist language teacher in the evaluation phase. They are meant to allow the user to access to some of the underlying information, in order to help them in their evaluation, adopting a qualitative point of view where prisms are quantitative.

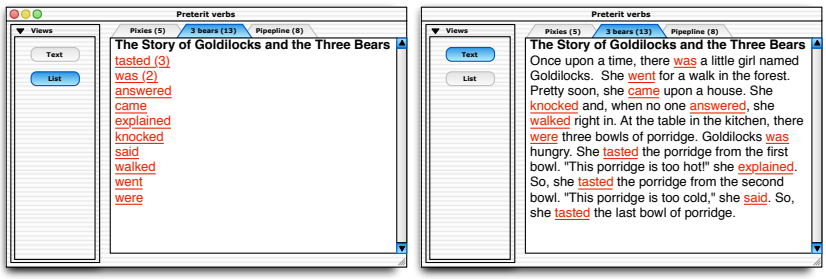

Fig. 6. Example of views linked to F_{repE^t} for PC *preterit*

For instance in figure 6, a user looking for a text to create a structural exercise on the preterit tense in English might want to make sure that, beyond

the number of occurrences, the text contains irregular verbs including "to be". To discard a text the list view would be sufficient and might be more convenient than the highlighted view (see figure 6). The latter would offer to the teacher an in context glimpse at the verbs, that might be preferable to make sure that the resulting activity would not prove too difficult (or easy) for the learners.

The notion of view has not been fully formalized yet. The link with facets has to be specified further: are some of the views completely independent from any facet (and thus prism), relying on their own pre-processing or should they all be linked to a facet the way the views in figure 6 are to F_{repE^t} ? Should the ones that are linked to specific facets solely be linked to them by their common pre-processing or should they before all be linked to a prism ?

6 Conclusion

We introduced this model as a second version of a previous work [11]. This new version is not only justified by a concern to make it clearer: despite being similar in philosophy, it comes after the theorization of the notion of Pedagogical Context. Even though present in the first version of the model, PC was roughly defined. The work on the notion has allowed us to build on sounder basis the notions of facet and prism, which have been subject to semantical alteration. The prism was in the first version a global module of the system handling all processes. It is now explicitly linked to a facet, thus underlying the tight link between the two of them.

Despite its simplicity, prism $P_{wrdCount}$ exemplifies this relation, the kind of approximation inherent to the task at hand and the usefulness of NLP in the implementation of such a system. Depending on the capacities of the pre-processing[13] the definition of the facet can be altered (or the other way around). The word count can be based on a list of separators between which lie the words to be counted. In this case, the French "*chou-fleur*" could be two words, while it actually designates a precise object (cauliflower)[14]. The decision of which kind of treatment to use can come from a didactic question: one wants to evaluate the length of the text, in order to provide an idea of size of the text, considering compounds as separate words might not be a problem. But one might consider that the word count should be as consistent with the linguistic definition of word as possible. Yet, what interests teachers could actually be to consider as words only non function words in order to get a better grasp at the quantity of vocabulary necessary to understand the text. On the other hand the choice of what the facet actually means might come from purely practical reasons: the available word count function works with no dictionary whatsoever and cannot distinguish function words from others or even identify a compound. In all

[13] In this case the pre-processing actually could evaluate the property, due to its independence from the PC.

[14] '-' should be a separator in French since it is added when the verb and subject are inverted to form a question: Dort-elle *? Oui,* elle dort *comme une masse.* (Is she sleeping? Yes she is sleeping like a log)

possible solutions, the link between the concept behind the facet and the prism should remain explicit and unaltered, might it mean modifying the prism, the facet or both...

The meaning of view has also changed (the view of this version of the model corresponds more or less to the visualization of the former) leading to alteration of the implementation. The questions raised in the previous section by this extension to the evaluation task are among the various implementation questions at hand. We have yet to implement a prototype of this version of the model. It will undoubtedly raise more questions, such as the storage of all the information added to documents, which can be partially answered by the many works in the field of multi-structured documents (see for instance [15]).

But one key issue in our opinion is that of the definition of a framework for prisms in order to make their integration and development easier. This issue could lead us to consider the problem of the system's adaptation to its users up to allowing them to create their own prisms and facets. Indeed we have seen with F_{AN} that a new prism could with didactic added value could be implemented with very little treatment (threshold values definition) beyond the grouping of two existing prisms. Careful analysis and specification of implementation consequences of the properties of prisms might constitute a viable path toward end-user programming functionalities [13] through the creation of compound prisms. This would not have been a perspective with the monolithic prism of the previous version of the model.

References

1. Antoniadis, G., Échinard, S., Kraif, O., Lebarbé, T., Ponton, C.: Modélisation de l'intégration de ressources TAL pour l'apprentissage des langues : la plateforme MIRTO. ALSIC 8(Numéro spécial TALAL), 65–79 (2005), http://alsic.u-strasbg.fr/v08/antoniadis/alsic_v08_04-rec4.htm
2. Balatsoukas, P., Morris, A., O'Brien, A.: Learning objects update: Review and critical approach to content aggregation. Journal of Educational Technology & Society 11(2), 119–130 (2008), http://www.ifets.info/journals/11_2/11.pdf
3. Bertrand, A., Cellier, J.M., Giroux, L.: Expertise and strategies for the identification of the main ideas in document indexing. Applied Cognitive Psychology 10(5), 419–433 (1996), http://www3.interscience.wiley.com/journal/21437/
4. edna: edna resources - metadata application profile (2006), http://www.edna.edu.au/edna/go/resources/metadata/pid/261
5. GEM: Listing of gem 2.0 top-level elements (2004), http://www.thegateway.org/about/documentation/metadataElements/index_html
6. Blanchard, A., Kraif, O., Ponton, C.: Mastering noise and silence in learner answers processing: simple techniques for analysis and diagnosis. CALICO Journal (2009), http://tr.im/calicoabokcp
7. Bourda, Y.: Des objets pédagogiques aux dossiers pédagogiques (via l'indexation). Document numérique 6(1-2), 115–128 (2002), http://www.cairn.info/revue-document-numerique-2002-1-page-115.htm
8. Charlier, É.: Planifier un cours, c'est prendre des décisions. Pédagogies en développement. Série 5, Nouvelles pratiques de formation, De Boeck Université, Bruxelles ; Paris (1989)

9. Final 1484.12.1 LOM draft standard document. Tech. rep., IEEE LTSC WG12 (2002), http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf
10. Loiseau, M.: Élaboration d'un modèle pour une base de textes indexée pédagogiquement pour l'enseignement des langues. Ph.D. thesis, Université Stendhal Grenoble 3 (2009), http://tel.archives-ouvertes.fr/tel-00440460/fr/
11. Loiseau, M., Antoniadis, G., Ponton, C.: Model for pedagogical indexation of texts for language teaching. In: Cordeiro, J., Shishkov, B., Ranchordas, A., Helfert, M. (eds.) ICSOFT (ISDM/ABF). vol. ISDM/ABF, pp. 212–217. INSTICC Press (2008), http://mathieu.loiseau.free.fr/bdtip/fichiers/articles/icsoft-2008.pdf
12. Loiseau, M., Antoniadis, G., Ponton, C.: Pratiques enseignantes et ≪ contexte pédagogique ≫ dans le cadre de l'indexation pédagogique de textes. In: Neveu, F., Muni Toke, V., Durand, J., Klinger, T., Mondada, L., Prévost, S. (eds.) 2^e Congrès Mondial de Linguistique Française. pp. 479–492. Institut de Linguistique Française, EDP Sciences, Paris (2010), http://www.linguistiquefrancaise.org/articles/cmlf/pdf/2010/01/cmlf2010_000233.pdf
13. Nardi, B.A.: A Small Matter of Programming : Perspectives On End User Computing. MIT Press, second printing (1995) edn. (1993)
14. Pernin, J.P.: Normes et standards pour la conception, la production et l'exploitation des EIAH. In: Grandbastien, M., Labat, J.M. (eds.) Environnements informatiques pour l'apprentissage humain, pp. 201–222. Hermès et Lavoisier, Paris (2006)
15. Portier, P.E., Calabretto, S.: Multi-structured documents and the emergence of annotation vocabularies. In: Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies (2010), http://dx.doi.org/10.4242/BalisageVol5.Portier01
16. Recker, M.M., Wiley, D.A.: A non-authoritative educational metadata ontology for filtering and recommending learning objects. Interactive Learning Environments 9(3), 255–271 (2001), http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=5848430&site=ehost-live
17. Scorm overview. Specifications SCORM 2004 3rd Edition Content Aggregation Model Version 1.0, ADL (2006), http://tr.im/scorm2004_3
18. Yinger, R.J.: A study of teacher planning: Description and a model of preactive decision making. Michigan State University, Institute for Research on Teaching, East Lansing, MI (1978), http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED152747

Appendix: acronyms

CALL	Computer Assisted Language Learning
edna	Educational Network of Australia
GEM	the Getaway to Educational Material
LOM	Learning Object Metadata
MIRTO	Multi-apprentissages Interactifs par des Recherches sur des Textes et l'Oral
NLP	Natural Language Processing
PC	Pedagogical Context
SCORM	Sharable Content Object Reference Model