The concept of "text facet" as a means to achieve pedagogical indexation of a text base dedicated to language teaching

Mathieu Loiseau, Georges Antoniadis, Claude Ponton

LAP-TaLC8

The concept of ``text facet'' as a means to achieve pedagogical indexation of a text base dedicated to language teaching ()

Author
Mathieu Loiseau and Georges Antoniadis and Claude Ponton

Booktitle
Proceedings of the 8^{th} Teaching and Language Corpora Conference (TaLC)

Year
2008

Editor
{Associação de Estudos e de Investigação Científica do ISLA-Lisboa}

Pages
421--425

Organization
ISLA

Publisher
Offsetmais Artes Gráficas, S.A.

Address
Lisboa

Colloque
TaLC 8

Lieu
Lisbonne

Dates
4-6 juillet 2008

Keywords
Corpus, Didactique des langues, ALAO

Url
http://mathieu.loiseau.free.fr/bdtip/fichiers/articles/Talc8.pdf

The concept of "text facet" as a means to achieve pedagogical indexation of a text base dedicated to language teaching

Mathieu Loiseau, Georges Antoniadis, Claude Ponton

LIDILEM, Université Stendhal Grenoble 3 – France

Abstract This communication is meant to present our project of pedagogically indexed text base. After introducing the notion of pedagogical indexation, which needs to be articulated around the teachers needs, we explain to which extent existing pedagogical resource description standards are inadequate to achieve pedagogical indexation for raw texts for language teaching. We then introduce the notions underlying the creation of our prototype through a fictional study case. The notions, which we introduce, are meant to be able to take into account the pedagogical context of the potential use of the text when giving a value to its pedagogical properties. These notions include text facet, view of a text according to a facet for a given pedagogical context, homogeneous text collection and text visualization. Through the use of these notions our prototype will allow teachers to query for texts depending on pedagogical criteria and provide them with assistance for the actual choice of the text.

Computer Assisted Language Learning (CALL), Pedagogical Indexation, Natural Language Processing (NLP), Pedagogical Resource Description

Introduction

Text base and pedagogical indexation

Despite the popularity of the communicative approach (Levy 1997:123) and the increased use of authentic texts[1], there is no text base available that allows teachers to query in language didactics relevant terms. Teachers have adapted some of their practices to existing computer tools, such as in Data Driven Learning (DDL) (Johns 1991) or proposed methodology for "pedagogic mediation of corpora" (Braun 2005).

[1] 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."

All the same, tool-wise, some flaws of CALL systems identified in (Antoniadis *et al.* 2004) remain representative of the situation of language corpora for language teaching: if a teacher seeks to find a text in a corpus, systems will not allow him/her to express his/her query in terms of his/her set of problems: using pedagogical concepts.

Our project of pedagogically indexed text base directly stems from the previous observation. As part of this project, a prototype is being implemented. In order to present some of the concepts underlying its design we define the notion of pedagogical indexation as "*indexation performed following a documentary language describing the objects according to pedagogical criteria (relevant to didactics)"*.(Loiseau *et al.* 2005).

Users' practices

In order to try to adapt the system to the actual teachers' practices, we decided to adopt an empirical approach: we performed a study in three parts. We initiated it with 8 interviews of language teachers of different experience, taught language and *computer literacy* (Bawden 2001). We used the information we gathered to prepare a short questionnaire destined to grasp how teachers handle authentic texts and the classification and research of texts. This questionnaire was answered by 133 teachers and allowed us to validate the hypothesis that a given text can be used in a variety of pedagogical contexts[2] (Loiseau *et al.* 2008). We also concluded that teachers favor authentic texts and that they resort to specially constructed texts when they want to control their linguistic content (grammatical structures, vocabulary), especially with beginners groups. We then issued a longer questionnaire, meant to precise the information gathered in the first questionnaire and to isolate research criteria.

421

Connecting thread

We will expose our conclusions, confront them with existing pedagogical resource description standards and introduce some of the concepts underlying the design of the prototype to the light of a virtual case study. We will imagine the case of an English

[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.)

teacher, whom we will call Bert for the sake of not repeating "the teacher / his or her / him or her" throughout the article. Bert wants to work with a group of students on the preterit tense and seeks texts in order to prepare some activities around this grammatical notion. We will follow Bert's steps throughout the article and try to show the actual consequences of the different modeling options on his practices.

Inadequacy of pedagogical resource description standards

In our virtual case study Bert wants to work on the preterit and therefore is most likely to be looking for a text containing occurrences of preterit[3]. Considering which is the best and most natural way for a teacher to phrase his/her query is a problem which ought to be addressed in a later version of the prototype[4], we will focus here on how to design the system so that it can answer this query: "text containing occurrences of preterit forms of verbs".

In order to retrieve resources based on their pedagogical properties various standards have been developed. Among these standards we are going to consider here the case of Learning Object Metadata (LOM) (IEEE 2002)., which is representative of all the standards we have studied[5]. LOM proposes a set of data elements (more than seventy) meant to describe the properties of a "pedagogical object" that is to say "*any entity – digital or non-digital – that may be used for learning, education or training*". A text to be used in a language learning activity undoubtedly satisfies this description.

Now, let us imagine Bert using a text base, the objects of which are described with LOM data elements. A text containing preterit verb forms can be described as "Text adequate for the introduction of the preterit tense" using LOM data element 5.10, "Description". But the text might also be adequate to perform a phonetic exercise on

---

[3] We assume here he does not look for a text to be used in a rephrasing structural exercise of the "Change the verbs in the following text into past tense" type.

[4] Preterite *vs* preterit *vs* simple past *vs* past historic *vs* (\w*ed) *vs* VPret, etc.

[5] Among which are Dublin Core Metadata Elements, GEM, EdNA, SCORM, IMS Metadata

compounds, to work on the lexical field of sports or any other use. Now, one can wonder whether it is possible to actually list every single potential uses of a text. From Bert's point of view, this means that a text described as adequate for work on the preterit tense will be so, but one that has not been described as adequate might be, all the same. The only way for Bert to figure out, is to actually read the articles, which cannot be considered a significant upgrade from his present practices.

Appropriately describing a text using LOM raises a concern of exhaustiveness, in that the index of the text would need to reference all the possible uses of the text. It is not because one annotator has considered the text fit for a given activity that it cannot be used for others. Indeed our second questionnaire not only confirms that a given text can be used in various pedagogical contexts, it also establishes that pedagogical properties of the text actually depend on one another. For instance the difficulty of the text (LOM descriptor 5.8, "difficulty") depends on what is to be made of it (LOM descriptor 5.10, "description") and with whom (LOM descriptor 5.7 "Typical age range"). LOM considers pedagogical properties as intrinsic to the resource described. In the case of raw resources, such as texts in the context of our work, an exhaustive description of the resource is bound to be extremely tedious, if feasible. We therefore need to focus on a different method of description, which would not require considering *a priori* all the possible combinations of properties of the text.

Text facets and views of a text according to a facet for a pedagogical context

Definitions

To be able to take into account the various parameters influencing the properties of the text we introduce the notion of text facet: "*a* text facet *is a property defined with a view to the text's pedagogical exploitation in language 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*".

422

Let us come back to Bert's query, looking for a text containing occurrences of preterit. A useful facet for this query would be a facet that we will call "representative elements

of a notion count". From now on, we will refer to this facet as F_{RepEt}. The mechanism involved to compute the value of this facet would regroup natural language processing (NLP) a morphological analyzer, a pattern matching program and a counter. Through his query, Bert specifies the pedagogical context by which to compute the value of the facet. In our case he wants to work on the preterit, the system will therefore compute for each text the number of representative elements of the notion "preterit" using the result of the morphological analysis. This value, computed for each text and depending on the pedagogical context is called a *view* of the text according to F_{RepEt} for the pedagogical context "preterit form of verbs".

Pedagogical context and constraints

With the views of the texts, Bert is now able to know whether each text contains occurrences of the preterit and how many. Given this information, he or she will obviously not be interested by certain texts (those not containing any occurrence for instance). Rather than letting him browse through all the texts contained in the text base, we could slightly modify F_{RepEt} in order to allow a more precise pedagogical context. Our study showed that depending on the kind of activity they want to perform with the text, teachers do not look for the same number of occurrences of the notion they wish to work on. Introducing the notion, for instance requires less occurrences than compiling a gap-filling structural exercise. It therefore seems relevant to let Bert constrain the value of F_{RepEt} through an extended pedagogical context changing his query to "texts containing at least 4 occurrences of preterit". This new constrained version of F_{RepEt} will be called F_{RepEtC}. The view of a text satisfying the condition according to F_{RepEtC} will return the number of occurrences of the structure. If the text does not satisfy the condition, its view according to F_{RepEtC} will be called empty. To a given pedagogical context, the system will yield all the texts with a corresponding non empty view.

Homogeneous text collection

At this point, Bert has been given access to a subset of texts satisfying his original query: "texts containing at least for occurrences of preterit". Still, depending on the number of text indexed and the variety of their content, he might be given a wide choice of texts. In such cases the teacher should be able to gradually refine his/her query using

different facets until the number of texts is low enough for him/her to choose one. This ultimate choice cannot be performed by the system, for it involves notions that are very difficult to model and compute automatically (such as the theme of the text) or that are not yet computable (the interest the students might have in the text).

In order to allow the user to gradually refine his/her queries, we introduce the notion of view of a collection of texts[6]. The view of a collection of texts C_1 according to a facet F for a pedagogical context CP is a collection C_2 of texts containing all the texts of C_1 the view of which is not empty (according to F for CP). We call C_2 a "*homogeneous text collection*" in that all the texts it contains at least have in common the property of satisfying the constraints enunciated in CP.

The consequence is that Bert does not necessarily have to formulate a complete query and can specify it further, provided that the system contains enough texts satisfying the most simple version of his query; and this, without having to compute the views all over again. To follow up on our example, we can imagine that Bert's query yielded too many results. In order to narrow down the choices we introduce a new facet: F_{WC}, which counts the number of words contained in the text. As we said it, Bert wants to introduce the notion based on a comprehension activity. He wants the text to be interesting, and does not require it to be too dense in occurrences of the preterit. His students are still beginners; the texts therefore need not be too long. He is looking for texts of 250 words (with a tolerance of 25%) among those containing 4 occurrences or more of preterit.

In figure 1 below, C_C stands for complete collection, C_1 could be the view of C_C according to F_{RepEtC} for the pedagogical context "at least for occurrences of preterit", as requested by Bert. In this case, C_2 is the view of C_1 according to F_{WC} for the pedagogical context "250 words ± 25%". We will explain in the next paragraph, the notion of visualization.

423

[6] We use the term "text collection" in order to distance ourselves from the constraints of corpora: "Words such as collection and archive refer to sets of texts that do not need to be selected or do not need to be ordered or the selection and/or ordering do not need to be on linguistic criteria. They are therefore quite unlike corpora." (Sinclair 1996)

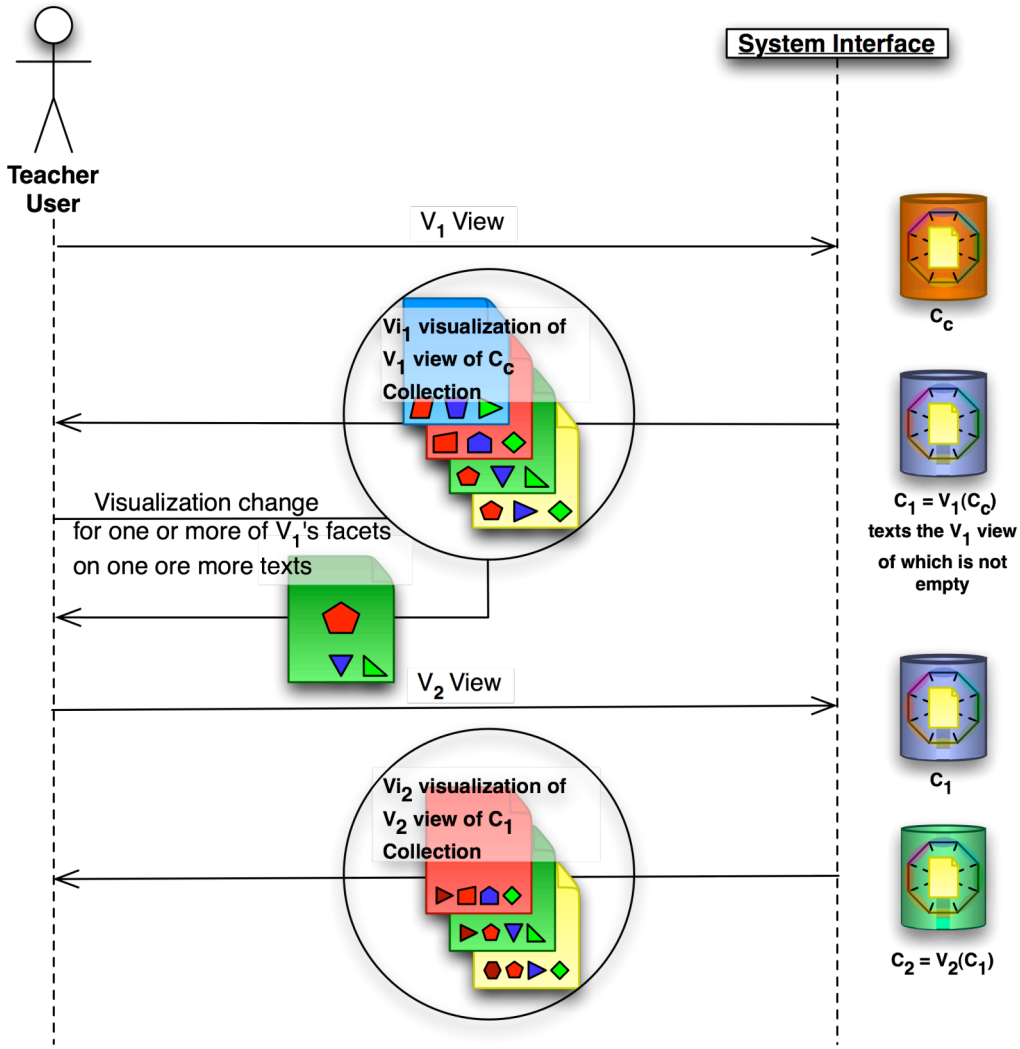

Figure 1 - Example of interaction between a language teacher and the system to refine his or her original query.

Qualitative access to the facets: visualizations

For the sake of our example, the system yields a dozen texts among which to choose. So far the system just acted as a filter based upon the pedagogical context submitted by the user. At this point, there are various criteria that the system as we have described it – a system implementing only two facets: F_{WC} and F_{RepEtC} – has used all the information it disposes of. The choice between the candidate texts can only be done by Bert himself, who knows what the center of interests of his students are, who can evaluate the adequacy of the text with the level of the students and what uses of the preterit he wants to display to his students. For the latter, the system is able to help: to be able to count the occurrences of preterit, it has annotated them and can re-use this annotation to offer

Bert some assistance. For each facet, the system associates one or more graphical representations, which we call visualization. These visualizations can either use the view itself or the underlying information used to compute it. For instance, in the case of F_{RepEtC}, computation of the view required a morphological analysis of the text. The system can reuse it to highlight all the occurrences of preterit or just present a list of all the preterit forms in the texts, instead of showing the whole text (cf. figure 2).

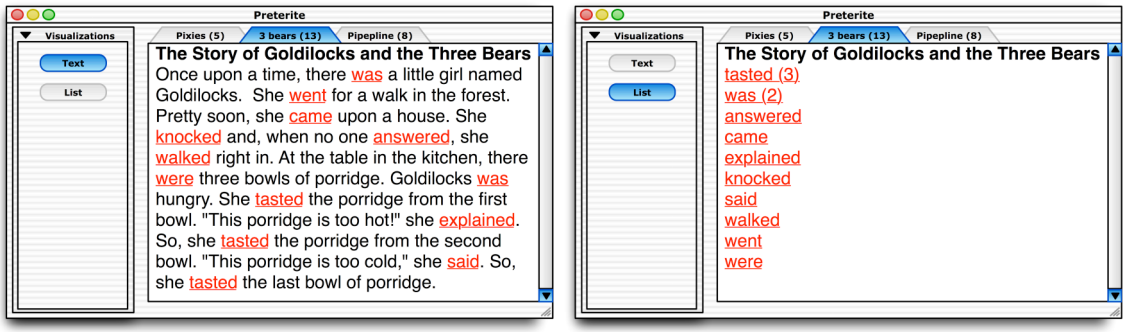

Figure 2 - Example of two different visualizations of the same text for the same facet

Conclusion

We have tried to explain, through a very simple example using very few tools, how the concepts of text facet, view and visualizations could help to acknowledge the influence of the pedagogical context on the pedagogical properties of the text. The pedagogical resource description standards while adapted to the description of already pedagogically exploited resources, do not seem fit to describe raw resources, in particular texts, in the context of their use in language teaching. The simplicity of the facets presented suggest that such a system could offer higher pedagogical added value, should other information or tools be made available to it. This is why we resorted to a very modular architecture, meant to regroup all the functions used in a treatment unit called the prism. Each facet is thus associated to a treatment sequence, using the functions grouped in the prism.

424

Combination of facets

This modularity is meant not only to be able to integrate to the prototype various sources of information, such as annotated corpora, or NLP tools, but also to reuse or combine existing facets. We believe that this architecture is evolvable enough to improve pedagogical indexation of text for language teaching through an iterative

process and the collaboration between teachers, didactics experts and computer scientists. Starting with the two very simple facets we have introduced here (F_{RepEtC} and F_{WC}), one can already start to create more evolved facets. Our study showed that the kind of activity the teachers meant to use the text in, influenced the number of representative elements of the notion at the center of the activity and on the length of the text. Based on our results, we could create a facet asking the teacher what he or she wants to do with the text and what kind of structures he or she is interested in confronting the students with. The system would then translate this information into threshold values and tolerance for F_{RepEtC} and F_{WC}. Of course, this new facet would rely on declared and not actual practices and would thus probably not be very powerful. Still the prototype could help gather information on actual practices to make the values more accurate. In turn, additional parameters could be taken into account, such as the level of the students, which also influences the length of texts depending on the kind of activity. The integration of the new parameters can, in the same way, be confronted to the teachers practices via the prototype to be fine tuned and so on. To be effective, iterative process should feed off research in language didactics, NLP and the conclusions which can be drawn from the use of the prototype.

References

Antoniadis, G., Échinard, S., Kraif, O., Lebarbé, T., Loiseau, M. and Ponton,C. 2004 "Nlp-based scripting for call activities." Paper presented at the *Coling Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*, Genève, August, 2004.

Bawden, D. 2001. "Information and digital literacies; a review of concepts" *Journal of documentation* 57/2: 218-259.

Braun, S. 2005. "From pedagogically relevant corpora to authentic language learning contents" *ReCALL* 17/1: 47-64.

I EEE LTSC WG1 2. 2002. *Final 1484.12.1 lom draft standard document*. http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf [Access date 15/02/2008]

Johns, T. 1991. "Should you be persuaded: two examples of data-driven learning" In *Classroom Concordancing*, T. Johns and P. King (eds) English Language Research Journal 4: 1-16.

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

Loiseau, M., Antoniadis, G. and Ponton, C. 2005. "Pedagogical text indexation and exploitation for language teaching " In *Recent research developments in learning technologies*, A. M. Vilas *et al.* (eds.). Badajoz: FORMATEX, 984-994.

Loiseau, M., Antoniadis, G. and Ponton, C. 2008. "Model for pedagogical indexation of texts for language teaching" Paper to be presented at the *3rd International Conference on Software and Data Technologies* (ICSOFT 2008), Porto, July 5 - 8 , 2008.

Sinclair, J. 1996. "Preliminary recommendations on corpus typology" *EAGLES (Expert Advisory on Language Engineering standards)* http://www.ilc.cnr.it/EAGLES96/pub/eagles/corpora/corpustyp.ps.gz [Access date 05/30/2008]

Taylor, D. 1994. " Inauthentic Authenticity or Authentic Inauthenticity?" *Teaching English as a Second or Foreign Language* ½: A-1.

The authors

Georges Antoniadis and Claude Ponton are both Maitre de conférence *at the LIDILEM laboratory in Grenoble. They have worked in NLP in automated generation of text and recently focused on the added value NLP functions could provide in CALL systems, in*

particular with the MIRTO platform. They supervise the PhD thesis of Mathieu Loiseau, the object of which is the creation of the model and prototype being discussed in this article.

```
@inproceedings{LAP-TaLC8,
    Author = {Mathieu Loiseau and Georges Antoniadis and Claude Ponton},
    Crossref = {:2008ol},
    Pages = {421--425},
    Title = {The concept of ``text facet'' as a means to achieve pedagogical indexation of a text base dedicated to language
teaching},
    Url = {http://mathieu.loiseau.free.fr/bdtip/fichiers/articles/Talc8.pdf}}

@book{TaLC:2008ol,
    Address = {Lisboa},
    B.U. = {DIP},
    Booktitle = {Proceedings of the $8^{th}$ Teaching and Language Corpora Conference (TaLC)},
    Colloque = {TaLC 8},
    Dates = {4-6 juillet 2008},
    Editor = {{Associa{\c c}{\~a}o de Estudos e de Investiga{\c c}{\~a}o Cient{\'\i}fica do ISLA-Lisboa}},
    Isbn = {978-989-95523-1-9},
    Key = {TALC},
    Keywords = {Corpus, Didactique des langues, ALAO},
    Language = {english},
    Lieu = {Lisbonne},
    Organization = {ISLA},
    Publisher = {Offsetmais Artes Gr{\'a}ficas, S.A.},
    Title = {Proceedings of the $8^{th}$ Teaching and Language Corpora Conference (TaLC)},
    Year = {2008}}
```