* GUILLERMO BARRUTIETA
** JOSEBA ABAITUA
*** JOSUKA DÍAZ

* Mondragon Unibertsitatea
Arrasate, Spain
gbarrutieta@eps.muni.es

** Universidad de Deusto
Bilbao, Spain
abaitua@fil.deusto.es

*** Universidad de Deusto
Bilbao, Spain
josuka@eside.deusto.es

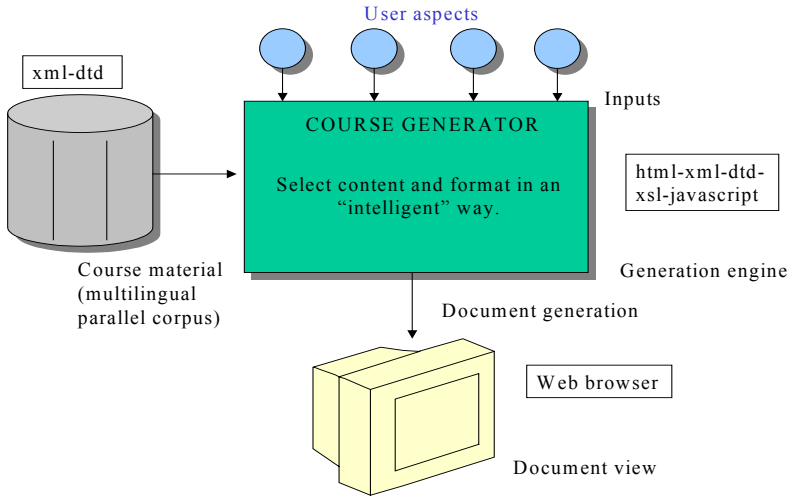
User modelling and content selection for multilingual document generation

Abstract

User modelling is a key aspect of content selection for successful document generation. This paper discusses how user aspects can be adequately ordered and parameterised following the results of a survey among professionals and students in the context of good document production. The model has been implemented in an experimental system for multilingual document generation in a learning environment (the CourseViewGenerator) that applies gross-grained RST to the elaboration of a master document which is later completed by a content selection algorithm.

Introduction

It is widely acknowledged that user modelling plays a crucial role in text generation (Paris 1987). This paper shows how user aspects have been modelled in a multilingual document generation system (the CourseViewGenerator) for educational purposes (Barrutieta 2001a and et al. 2001b). The general objective of this system is to generate multilingual learning documents automatically. These documents have to be adequate to the students



General schema of the multilingual document generation system

The following two aspects must be explained:

1. How a good document may be defined.
2. What we do know about the users of those documents and their needs.

A good document is a readable and easy to understand document, but mainly a good document should meet the user needs. We are going to concentrate on this last attribute of a good document since our approach is a user focused approach and not text focused or expert focused approaches in a three-fold distinction developed by Schriver (1989).

Our users (readers) are university students of Mondragon Unibertsitatea, at least 18 years old, both male and female, intellectually fit (no students with learning disabilities) and living in the Basque Country or the nearby provinces. They are therefore likely to be bilingual in Basque and Spanish and have sufficient English reading skills.

Students vary in their interests, in their needs and in their capabilities. Ideally, therefore, each student on a course should receive personalised textual material, whose design and layout suit that student as an individual (Lefrere 1995).

But also at the start of the learning process, students need to be able to browse and to travel through the materials. At a later stage they may want to approach the materials from an evaluation point of view and construct study modes to work through the material. This is why we believe it is essential, as soon as it is practicable, to provide students with ways of modifying their learning material to suit their evolving needs. (Lefrere 1995).

We think that natural language generation techniques and user modelling techniques must play a paramount role in the personalised learning material of the future.

In the following sections of this paper, we present the results of the inquiry we conducted to find out about what an author takes into account when producing documents that try to meet the user needs. We show the user aspects that we found, not only the most frequent but also the most important user aspects (according to our research), and we also show how these aspects were implemented in the *CourseViewGenerator*.

What we asked

The following question was prepared in order to find out from a set of professionals and students what they take into account when they produce documents that must be relevant to the users they write for.

> Imagine that you are requested to put together a document about a certain subject. You want the document to be as relevant and easy to understand as possible for whoever has to read it. You do not know the readers but I do. What would you ask me about the readers?

Who we asked

We asked the question above to 25 professionals with a university degree currently working in all sorts of corporations, research and educational institutions. They write documents regularly as part of their jobs.

We also asked the same question to 20 students that are currently in their 4th academic year at our university. They too have to write documents regularly in order to do their academic assignments.

First results: frequency table

In the following table we display the aspects that appeared when we collected the answers from the questionees. The first column contains the user aspect and the second column contains the frequency in which the aspect was given as an answer. This table contains the results from both the students and professionals, 45 people in total.

User Aspect	Frequency
1. Knowledge about the subject matter before reading the document	34
2. Time available to read the document	29
3. Reason to read the document	27
4. Age	19
5. Education and education level	18
6. Language and nationality	17
7. Social, economic and cultural situation and level	12
8. Preference of text versus images	12
9. Preference towards access structures of text (indexes, …)	8
10. Number of readers and diversity	6
11. Preference towards rhetorical structures	5
12. Job	5
13. Interest in other subjects (other than the subject of the document)	5
14. Location when the document is read	5
15. Opinion about the subject (if any)	5
16. Preferences towards bibliographical references and links to other documents	5
17. Gender	5
18. Relation to the subject of the document	4
19. Personal situation (busy, tired, …)	1

Some of the aspects that were collected had to be interpreted by us because different people refer to the same aspect differently although it seemed to us that they really meant the same thing.

We thought that a frequent aspect was probably important; in other words, we thought that a frequent aspect was in fact taken into account by the author but we wanted to be reassured on this. With this purpose we asked 9 of the professionals to order the 12 most frequent aspects in order of importance beginning by the most important aspect to be taken into account and ending by the least important aspect.

We gave them the following 12 aspects to reorder in terms of importance: 1, 2, 3, 4, 5, 6, 7, 10, 12, 13 15 and 17.

Some of the aspects were removed from the list because we understood that these aspects were related to the document and not to the user: 8, 9, 11 and 16.

And some other aspects were removed because we understood that the information was sufficiently conveyed by the 12 chosen aspects or we thought they were vague or irrelevant for our work: 18, 19 and 14.

The list of 12 aspects given to the questionees was ordered in terms of frequency (but this detail was not disclosed to them).

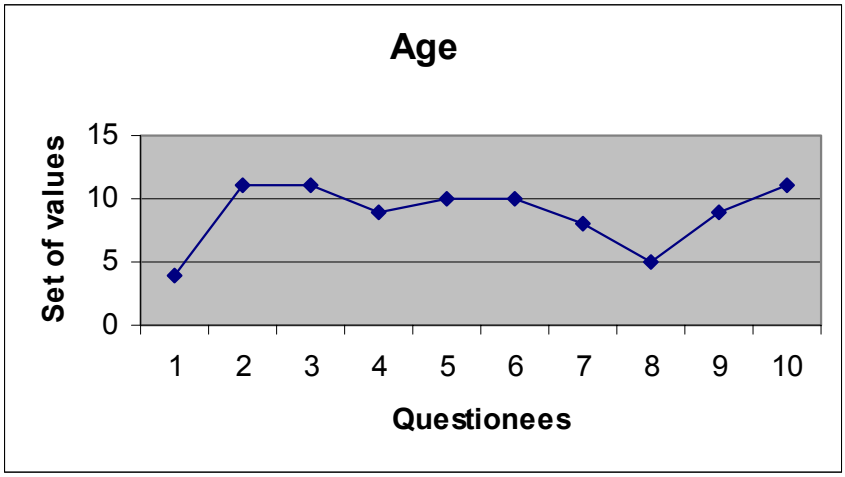
Second results: ordering table and graphs

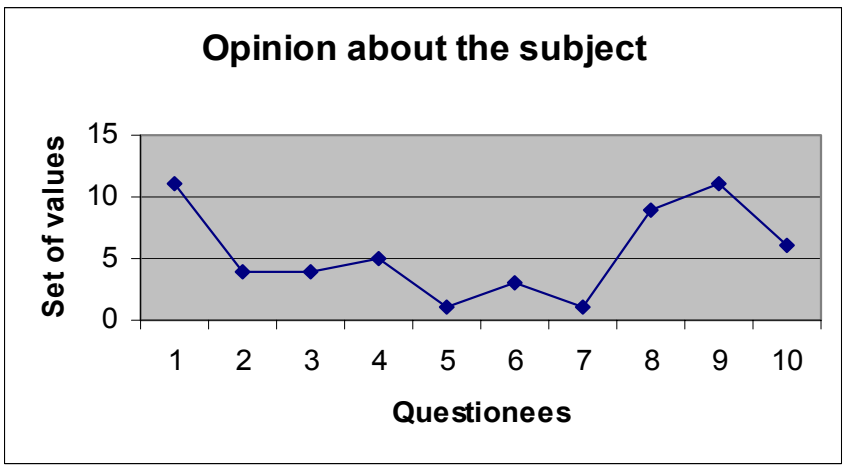
When we got the results of the ordering we did not encounter any big surprises. Most of the most frequent aspects were considered also important by the professionals that took part.

The following table contains the user aspect in one column and the average position value in the other column. The average value was calculated from the different values in the lists ordered by importance by the professionals:

User Aspect	Average position
1. Knowledge about the subject matter before reading the document	2
2. Time available to read the document	4.89
3. Reason to read the document	3.56
4. Age	9.3
5. Education and education level	4.6
6. Language and nationality	8
7. Social, economic and cultural situation and level	7.4
8. Number of readers and diversity	7.7
9. Job	6
10. Interest in other subjects (other than the subject of the document)	9
11. Opinion about the subject (if any)	4.9
12. Gender	11

However we did find two interesting results: “age” was frequent but so important and “opinion about the subject” was not frequent but it was considered important. The following two graphs show the results of this two aspects:





In these two graphs the first point is the frequency position and the following points are the importance position decided by the questionees.

Examining the values of the ordering table we chose the 5 most important user aspects: 1, 3, 5, 11 and 2.

It is interesting to point out that these 5 aspects are the 5 most frequent aspects except for the "age" (4) that disappears from the top list and the "opinion about the subject" (11) that joins the top 5 list from a low position in the frequency table.

User modelling in user aspects - implementation

Find below the implemented list of user aspects with the discrete possible values that can be selected for each aspect. We have the 5 general user aspects that were identified in the process described so far and 3 added domain specific user aspects.

Specific User Aspects	**Discrete values**
Subject	Language processors
Moment in time	Before the course/ Day 1/ Day 2/ … / After the course (review)
Languages	EN/ ES/ EU
General User Aspects	**Discrete values**
Level of expertise	Null/ Basic/ Medium/ High
Reason to read	To get an idea/ To get deep into it
Job or studies	Not related to the subject/ Related to the subject
Opinion or motivation	Against/ Without an opinion or motivation/ In favour
Time available	A little bit of time/ Quite some time/ Enough time

Discussion: connections between gross-grained RST and the user aspects

The work that we are currently conducting involves finding the connections between the gross-grained RST (Barrutieta et al. 2001) and the user aspects & their values presented here. These connections will be implemented in the content selection algorithm. Basically the selection algorithm has to choose the parts of a master document (Hirst et al. 1997) that are relevant for the user needs described by each value of each user aspect.

For example, the algorithm will discriminate satellites like elaboration and background if the user has a little bit of time to read the document. On the other hand, the algorithm will select satellites like example if he/she has a basic level of expertise because the examples will help him/her understand the subject matter. The content selection algorithm will also have to negotiate apparent contradictions such as the user has a little bit of time and wants to get deep into the subject assigning weights that make certain aspects more prevalent in certain circumstances.

Acknowledgements

Our thanks go to all of those anonymous people who took time from their busy agendas to kindly answer the question or request that we asked them to do for this research work.

References

Barrutieta G. (2001a) Generador inteligente de documentos de formación. Virtual Educa 2001, Madrid (Spain).

Barrutieta G., Abaitua J. & Díaz J. (2001b) Gross-grained RST through XML metadata for multilingual document generation. MT Summit VIII (IAMT-EAMT), Santiago de Compostela (Spain).

Paris C.L. (1987) The Use of Explicit User Models in Text Generation: Tailoring to a User's Level of Expertise, PhD Thesis, Columbia University.

Schriver K.A. (1989) Evaluating text quality: the continuum from text-focused to reader-focused methods IEEE Transactions on Professional Communication 32(4), pp. 238-55.

Lefrere P. (1995) Electronic layout and design visions of the future. Open and distance learning today. Routledge, London (United Kingdom).

Hirst, G., DiMarco, C., Hovy E., Parsons K. (1997) Authoring and Generating Heath-Education Documents That Are Tailored to the Needs of the Individual Patient. Proceedings of the Sixth International Conference. UM97. Vienna (NY-USA).