Integrated generation of graphics and text: a corpus study

Marc Corio and Guy Lapalme
<corio@IRO.UMontreal.CA>
<lapalme@IRO.UMontreal.CA>
Département d'informatique et de recherche opérationnelle
Université de Montréal, CP 6128, Succ Centre-Ville
Montréal Québec Canada, H3C 3J7

Abstract

We describe the results of a corpus study of more than 400 text excerpts that accompany graphics. We show that text and graphics play complementary roles in transmitting information from the writer to the reader and derive some observations for the automatic generation of texts associated with graphics.

For the past few years, we have studied the automatic generation of graphics from statistical data in the context of the PostGraphe system (Fasciano, 1996; Fasciano and Lapalme, 1998) based on the study of graphic principles from such diverse sources as Bertin (1983), Cleveland (1980) and Zelazny (1989). PostGraphe is given the data in tabular form as might be found in a spreadsheet; also input is a declaration of the types of values in the columns of the table. The user then indicates the intentions to be conveyed in the graphics (e.g. compare two variables or show the evolution of a set of variables) and the system generates a report in LaTeX with the appropriate PostScript graphic files. PostGraphe also generates an accompanying text following a few simple text schemas. But before adding new schemas, we have decided to make a corpus study of texts associated with graphics and this paper presents the results of this study. We studied more than 400 texts and we will show that the saying "a picture is worth a thousand words" needs to be modulated because graphics and text are far from being interchangeable and that their interactions are quite subtle. With hindsight, this may seem obvious but, without a corpus study, we could not have documented this result. Although multimedia systems have been studied for many years, we are not aware of any previous corpus study of the same scale.

1 Overview of PostGraphe

Many sophisticated tools can be used to build a presentation using statistical graphs. However, most of them focus on producing professional-looking graphics without trying to help the user to organize the presentation. To help in this aspect, we have built PostGraphe which generates a report integrating graphics and text from a set of writer's intentions.

The writer's intentions can be classified according to two basic criteria: structural differences and contents differences. We refer to intentions derived from structural differences as **objective intentions** and intentions derived from contents differences as **subjective intentions**. This definition stems from the fact that when differences between two intentions are more content than structure related, the writer is choosing what to say and not how to say it. The writer is thus making a subjective choice as to what is more important.

In our research, we have built a classification of messages, given in figure 1, based on Zelazny's (1989) work. At the first level, our classification contains 5 categories two of which have sub-categories obtained by using a fractional modifier.

For comparison, the fractional modifier indicates that the comparison should be done on fractions of the whole instead of the actual values. For distribution, we obtain a specialized intention where the classes are presented according to their fraction of the total. At the second level, the intentions become specialized according to subjective criteria.

These simple intentions can then combined either by composition or superposition. In composition, the order of the variables is important and there is a dominant intention; for example, the comparison of evolutions is quite different

Objective Structure *How to say ?*	Subjective Content *What to say ?*
Reading(V)	
Comparison(S_1,S_2) Comparison Fractional(V,S)	
Evolution(V_1,V_2)	Increase Decrease Stability Recapitulative
Correlation(V_1,V_2)	
Distribution(V,S) Distribution Fractional(S)	

Figure 1: Two level decomposition of simple intentions: V is a variable and S is a set of variables

from the evolution of a comparison. For example, *Sales figures of Xyz increased less quickly than the ones of Pqr between 1992 and 1994* compares evolutions while *Pqr always stayed at the top except between 1992 and 1994* shows the evolution of the comparison. In superposition, the intentions are merely expressed using the same graphic but the intentions do not interfere.

Figure 2 shows the the part of the Prolog input specifying the intentions and the output from PostGraphe. The intentions are divided in 2 sections: the first presents the 3 variables (*year*, *company* and *profits*). The second presents the comparison of the profits between companies and the evolution of the profits along the years.

We have also "ported" this idea of taking account of the writer's intentions into the spreadsheet world by creating an alternative *Chart Wizard* for *Microsoft Excel* which asks for the intentions of the user (comparison, evolution, distribution ...) instead of prompting for the sort of graphic (bar chart, pie chart ...); see (Fasciano and Lapalme, 1998) for more information.

2 Text and graphics integration

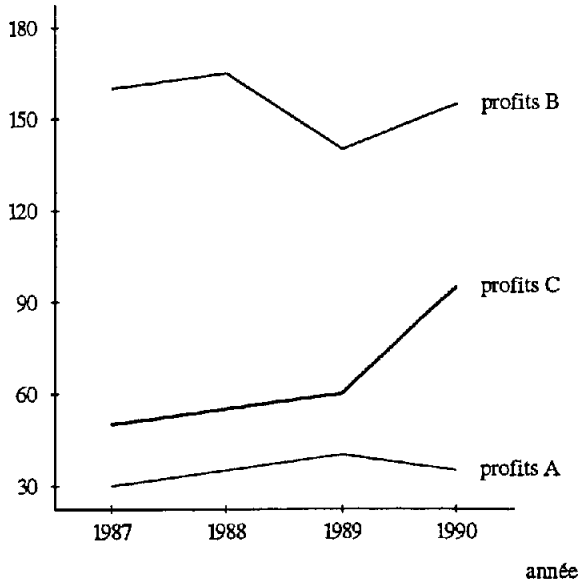
Graphics and text are very different media. Fortunately, when their integration is successful, they complement each other very well: a picture shows whereas a text describes. To create an

```
data(...
     % the intentions
      [[lecture(année), lecture(compagnie),
        lecture(profits)],
      [comparaison([profits],[compagnie]),
       evolution(profits,année)]],
     % the raw data
      [[1987,'A',30],...]).
```

Nouvelle section (3 intentions à traiter).

année	1987	1988	1989	1990
compagnie	profits	profits	profits	profits
A	30	35	40	35
B	160	165	140	155
C	50	55	60	95

Nouvelle section (2 intentions à traiter).



De 1987 à 1989 les profits de la compagnie A ont augmenté de 30 $ à 40 $. Jusqu'en 1990 ils ont diminué de 40 $ à 35 $.
De 1987 à 1988 les profits de B ont augmenté de 160 $ à 165 $. Pendant 1 année ils ont diminué de 25 $. Jusqu'en 1990 ils ont augmenté de 140 $ à 155 $.
De 1987 à 1990 les profits de C ont augmenté de 50 $ à 95 $.

Figure 2: Input specifying the intentions and the French Report generated by PostGraphe

efficient report from tabular data, choices must be made between modes of expression (text vs graphics) by taking into account their strong points but also their effect on the reader.

Graphics are usually floating elements that do not appear just beside the accompanying text, they are often moved to satisfy other graphical constraints such as avoiding blank space at the end of a page. Graphics make important elements of the data standout and catch the eye of the reader. The text of the report does not only describe or analyse data but it also links with the graphics by means of references to reinforce the intentions of the writer. Text and graphic coordination pose important problems identified in (Roth et al., 1991) as

structural incompatibility: text and graphics do not compose in the same way: for example, in a graphic representation of a tree, dependents are near the root but in a preorder textual description, the links might be harder to make;

lack of cohesion: for a text to make an explicit link with the graphical elements of an illustration, the text generator must have access to the structural elements of the graphics;

redundancy: the text should not repeat information that is better given by the graphics, although in a few cases it is a good idea to re-emphasize important information in the text.

3 Corpus study

As we want to generate not only well formed text but appropriate ones that complement the information available from the graphics, we have built a corpus of 411 French texts associated with graphics from such diverse sources as "Tendances sociales" published every three months by Statistics Canada, books on statistics, investment funds reports, governmental reports, etc.; see (Corio, 1998) for details. Like with most corpus studies, it is very hard to affirm that this study is representative but we have tried not to bias the kind of texts in any way except for cases when we detected that either the text or the graphics were not appropriate given the principles alluded to in section 2.

The analysis of our corpus revealed 7 main themes for texts combined with graphics. Table 1 gives the frequencies of each theme for the intentions described in figure 2. We now briefly describe each theme with a few examples. Finally, we will raise some automatic text generation issues that were the main motivations for this study.

descriptive gives an overview of the graphic or identifies its main visual aspect: for example, using a title or a legend, it describes the data on the X or Y axis or the general tendency (increase or decrease). Often this description identifies a selection criteria for the data such as *Ten OCDE countries having the highest percentage of adults registered to a University* which indicate that the graphics only gives a partial view of the data.

This theme is mainly associated with reading (73%) and evolution intentions (22%).

quantitative messages select the raw data that should interest the reader because, for example, the reader is directly concerned with this value: for a bar chart giving the annual income of a group of cities *The annual income of a Vancouver family was 59 700$ in 1993* is particularly interesting for somebody who lives in the Vancouver area or if it illustrates an article that deals with Vancouver.

It is interesting to see that many quantitative messages of our corpus refer to data that do not appear in the graphics; for example, the graphics shows a pie chart giving a budget distribution for 1997 but the text compares those figures with the ones of the previous year.

This theme is mainly associated with comparison (46%), evolution (30%) and reading (23%) but it is almost always possible to generate a quantitative message from any data either as it is or after some transformation such as a mean, a sum or by giving the range of the values.

domination expresses the highest or lowest values of the data such as *Which company made the most or the least profit.* Our corpus shows that sometimes the 2 or 3 dominating values are identified when these are

	reading	comparison	evolution	correlation	distribution	total	%
Descriptive	98	6	30			134	30
Quantitative	23	46	30			99	22
Domination		65	3		17	85	19
Deductive	11	16	5	33	3	68	15
Discriminant		7	31			38	9
Qualitative		2	8	3		13	3
Justificative		4		1		5	1
Total	132	146	107	37	20	442	100
%	30	33	24	8	5	100	

Table 1: Counts of themes and intentions of messages in our corpus of 411 French texts; some texts carry more than one intentions and theme

clearly separated from the rest. The messages can also indicate if the dominating values are for all possible cases. *In Canada, adults in the Newfoundland do the least sport* can only be said if all provinces are shown on the graphics.

This theme is associated with comparison (76%) or distribution (20%) intentions; in the case of a fractional modifier, the dominating values are in terms of percentages but for distribution, domination is indicated by an interval instead of specific data.

deductive messages draw a conclusion from the shape of the graphics or the values of the data; it can be either some form of correlation, a characteristic or a constant value in the data. These messages often use extra information to draw some conclusion. For example, *Provinces of western Canada had the highest employment rate for teenagers in 1993* makes use of geographic knowledge to link seemingly unconnected data: British Columbia, Alberta, Saskatchewan are part of western Canada but that fact is not explicitly given in the data for each of the ten provinces.

This theme is not closely linked with any particular intentions although correlation (49%) and comparison (24%) occur most often.

discriminant messages identify a particular fact that distinguish this value from the others: we show an irregularity, a turning point in a curve is identified or an exception in an otherwise constant situation.

This theme is associated with evolution (82%) and comparison (18%) intentions.

qualitative messages describe data in words such as *rare, weak, strong, frequent, high, low*; the shape of the curve can also be given. Here the judgement of the writer has the highest influence because the same value can qualified differently depending on the context.

These messages are most often associated with evolution (62%) intentions but they can also be encountered with correlations (23%) and comparisons (15%).

justificative messages identify causes for phenomena such as *Why is a bar the highest?, Why the canadian dollar fell?, Why a given political party has more voting intentions?.*

As our corpus has been mostly built from small texts we do not have enough data to associate this theme with particular intentions. These kinds of messages are most often met in longer texts.

3.1 Text and graphics interaction

It is often thought and said in the multimedia generation folklore and in some graphic generation texts that to obtain a good interaction between text and graphics, that text should give informations that the graphics does not show. But in our corpus, we observed most often that the text merely reinforces what is already evident in the graphics. For example, 29% of texts associated with a comparison intention, there is a mention of the highest value as to say to the reader: "Yes, what you see in this graph-

ics is really what is important". Redundancy only occurs when the text repeats exhaustively *all* the information and not when it pinpoints some important facts already "obvious" in the graphics.

Cohesion between text and graphics does not depend mainly on the type of graphics (bar chart, pie chart, etc.) but more on the type of data on each axis. For example, in a graphic illustrating the sentence *There are more graduates in the highest salary brackets*, data might be represented in salary intervals that can either be shown as bars, as columns, as an area under a curve or even as pie pieces. Thus each type of data has its own lexicon to insure cohesion: *tendencies* and *evolution* refer to a temporal axis no matter if the graphics is a curve or a bar chart.

In our corpus, there are few coreferences to visual elements of the graphics, but we believe that this phenomenon is specific to our domain of statistical data. We are quite sure that in the domain of instructional texts, references to graphical elements occur more often.

3.2 Lessons learned for automatic generation

From this corpus study, we developped some rules for selecting appropriate comments associated with the graphics chosen by PostGraphe while not overburdening the user with special annotations for the data. But as we saw that the texts are used to pinpoint some important aspects of the data, we need to know the interests of the user in much the same way as PostGraphe needs to know the intentions of the user like the Vancouver example given in the previous section. The system must also know if a set of nominal values form a complete enumeration to affirm that a value is the *lowest ranking* or if it deals with *the ten most important countries.* There is also the problem of knowing if it is appropriate to mention the crossing point of two curve or not or to speak about the reversing of a tendency.

Data must also be identified with sufficient detail to be described in the text. The system cannot infer that a given percentage is the *rate of persons charged of impaired driving* without being given explicitly.

The system must also be aware of the appropriate vocabulary to qualify certain types of data. For example 5% might be qualified as *low* for certain income tax rate but might be thought as *high* if it deals with an inflation rate in North America these days.

Messages that draw a general conclusion such as *Canadian families have been quick to adopt new information technologies in their home* are quite difficult to generate automatically. The same can be said of justifications or links with the outside world such as those found in stock market reports (Kukich, 1983). For example, it is impossible to generate *The price of gold dropped because of the BRE-X scandal* from the raw data of transactions on gold.

For our text generation module, we will thus need a few more informations from the user such as the list of variables that are more important to the writer and a slightly more explicit naming of the variables. As these informations are of utmost importance for the writer, they should not be a burden to find and give. If they are, then that means that the intentions of the writer are not clear.

4 Conclusion

Our system is not the first one to combine text and graphics (see for example, multimedia generation systems like COMET (Feiner and McKeown, 1991), SAGE (Roth et al., 1991) or WIP(André et al., 1993)). In our case, the output looks much simpler but our corpus analysis shows that, even in this case, the text generation concepts necessary to combine with these seemingly simple graphics is quite involved because it must rely on the intentions of the writer which are often left implicit. Even when they are given, complexity comes from the combinations of both media and intentions.

Acknowledgments

We thanks Massimo Fasciano for fruitful discussion about his work and his collaboration on this project. This project has been partially funded by a student grant from FCAR (Gouvernement du Québec) and a research grant from NSERC (Governement of Canada).

References

E. André, W. Finkler, W. Graf, T. Rist, A. Schauder, and W. Wahlster. 1993. WIP: the automatic synthesis of multimodal pre-

sentations. In M. T. Maybury, editor, *Intelligent Multimedia Interfaces*, pages 75 – 93. AAAI Press, Cambridge, MA.

Jacques Bertin. 1983. *Semiology of Graphics*. The University of Wisconsin Press. Translated by William J. Berg.

William S. Cleveland. 1980. *The Elements of Graphing Data*. Wadsworth Advanced Books and Software.

Marc Corio. 1998. Sélection de l'information pour la génération de texte associé à un graphique statistique. Master's thesis, Université de Montréal.

M. Fasciano and G. Lapalme. 1998. Intentions in the coordinated generation of graphics and text from tabular data. *submitted to Natural Language Engineering*, page 27p., January.

Massimo Fasciano. 1996. *Génération intégrée de textes et de graphiques statistiques*. Ph.D. thesis, Université de Montréal.

S. Feiner and K. McKeown. 1991. Automating the generation of coordinated multimedia explanations. *Multimedia Information Systems*, 24(10):33–41, October.

Karen Kukich. 1983. Knowledge-based report generation: A technique for automatically generating natural language reports from databases. In *Proceedings of the ACM SIGIR Meeting*, pages 246–250. ACM.

Steven F. Roth, Joe Mattis, and Xavier Mesnard. 1991. Graphics and natural language as components of automatic explanation. In Joseph W. Sullivan and Sherman W. Tyler, editors, *Intelligent User Interfaces*, Frontier Series, chapter 10. ACM Press.

Gene Zelazny. 1989. *Dites-le avec des graphiques*. InterÉditions.