Generation of texts for information graphics

Marc Corio
Guy Lapalme
Département d'informatique et de recherche opérationnelle
Université de Montréal, CP 6128, Succ Centre-Ville
Montréal Québec Canada, H3C 3J7
{corio,lapalme}@iro.umontreal.ca

Abstract

We describe SelTex a text generation system for producing short texts and captions to accompany information graphics that are generated according to the writer's intentions. SelTex uses rules that were extracted from a corpus study of more than 400 text excerpts. This corpus study shows that text and graphics play complementary roles in transmitting information from the writer to the reader. We then derive some observations for the automatic generation of texts associated with graphics many of which were implemented in SelTex.

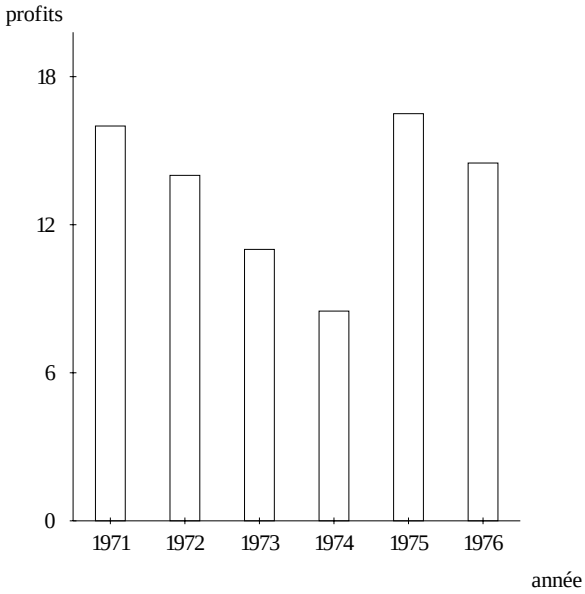
For the past few years, we have studied the automatic generation of graphics from statistical data in the context of the PostGraphe system [5, 6]. PostGraphe is given data in tabular form as might be found in a spreadsheet; also input is a declaration of the types of values in the columns of the table. 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 for only a few simple text schemas. Although PostGraphe has shown the potential of graphic generation, its text generation capabilities are well below its level of graphic competence. So we decided to redesign the text generation module to better capture the subtle nature of the interaction between text and graphics. Before designing SelTex, the new text module to be integrated in PostGraphe, we did a corpus study of more than 400 texts associated with graphics. The preliminary findings were presented in [3]. They will be reviewed here briefly before describing the text generation rules that were deduced from this corpus study. We will focus on the selection criteria for relevant facts that are expressed in the text that accompanies a graphic.

1 Overview of PostGraphe

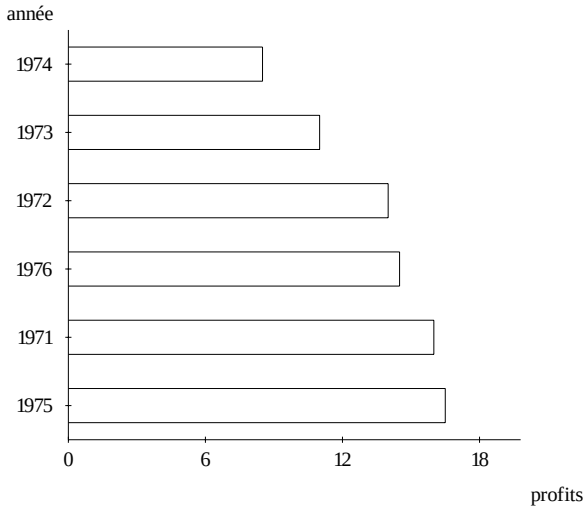
Many sophisticated tools can be used to create presentations using statistical graphs. However, most of them focus on producing professional-looking graphics without trying to help the user to organize the presentation. To help in this aspect, we have built PostGraphe which generates a report integrating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

In figure 1, the writer's intentions have an important effect on the way information is presented. Indeed, the figure shows the same data presented in two different ways according to what the writer wants to convey. One part of the example shows how to present the evolution of the data and the other part shows how to compare its various elements.

The writer's intentions can be classified according to two basic criteria: structural differences and contents differences. Figure 1 illustrates the structural difference between evolution and comparison. We refer



Globally, the profits have gone down despite a strong rise from 1974 to 1975.



The profits were at their highest in 1975 and 1971. They were at their lowest in 1974, with about half their 1975 value.

Figure 1: Two communicative goals: evolution and comparison

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 the difference between two intentions is more contents-related 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 Table 1, based on Zelazny's [11] 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. We have only studied subjective variations of evolution. In the first 3, the data is presented in a way that emphasizes the desired property. For example, it is possible to present only periods of increase from a data set or to group the data in longer intervals to lessen the effect of variations. The last subjective intention is a rather unconventional way of presenting temporal data: a non-temporal evolution. Although strange at first glance, this method is very effective in a text and takes a more global approach to evolution, trying to summarize long periods non-sequentially. For example, one might say that during a given period, profits have increased during two fiscal years and decreased during five without saying when and by how much. All of these subjective intentions, especially the last, could be criticised as being dishonest but this is how people prepare real reports.

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 from the evolution of a comparison like in the following sentences: *Sales figures of Xyz increased less quickly than the ones of Pqr between 1992 and 1994* compares evolutions while

Objective Structure *How to say ?*	Subjective Content *What to say ?*
Presentation(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)	

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

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 with the SelTex text generation module. The intention of the graphic is comparing the use of computers in each province of Canada.

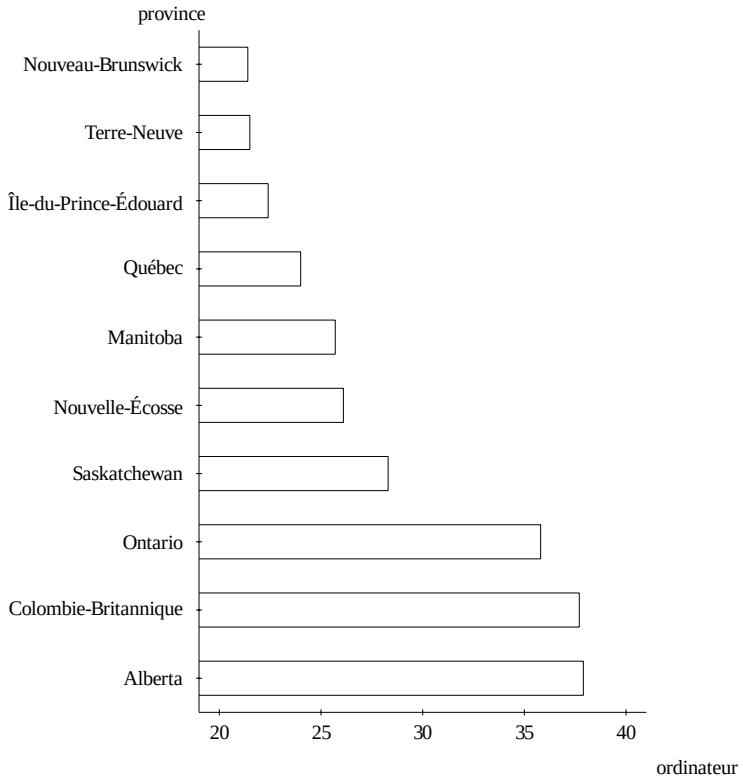
2 Corpus study

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. Like with most corpus studies, it is very hard to state that this study is representative but we have tried not to bias the kind of texts in any way. We kept all texts associated with a graphic that could have been generated by PostGraphe but we rejected some generic titles giving no intentional message such as *Mean familly income in 1995 constant dollars*; it is a shame to observe that these types of captions occur quite often in statistical reports. See [4] for more details on the process of collecting the corpus.

For each text excerpt, we assigned a classification code indicating the kind of information that is transmitted and the way it is expressed. These codes were revised many times but finally we identified 7 main themes for texts combined with graphics. Table 2 gives the frequencies of each theme for the intentions described in table 1. We now briefly describe each theme with a few examples. We then 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 presentation (73%) and evolution intentions (22%).

```
data([province,ordinateur], % variable names
                            % description of variables
     [province,'menages ayant un ordinateur'],
                                % types of variables
     [province,pourcentage],
     [province,taux],           % description of types of variables
     ['Quebec'],                % data of special interest
     [province],                % variables that can be relational keys
     [ordinateur],              % variables that should not be rel keys
                                % writer's intentions
     [[comparaison([ordinateur],[province])]],
     % data values
     [['Terre-Neuve',          21.5], ['Ile-du-Prince-Edouard',22.4],
      ['Nouvelle-Ecosse',      26.1], ['Nouveau-Brunswick',    21.4],
      ['Quebec',               24.0], ['Ontario',              35.8],
      ['Manitoba',             25.7], ['Saskatchewan',         28.3],
      ['Alberta',              37.9], ['Colombie-Britannique', 37.7]]).
```



L'Alberta, la Colombie-Britannique et l'Ontario ont un taux plus élevé de ménages ayant un ordinateur, tandis que le Québec se situe au septième rang avec 24.0 %. (*Alberta, British-Columbia and Ontario have a higher rate of households with a computer, while Québec is seventh with 24%.*)

Figure 2: Input specifying the intentions, the graphic chosen by PostGraphe and the the French text generated by SelTex followed by its English translation (done manually).

	presentation	comparison	evolution	correlation	distribution	total	%
Descriptive	98	6	30			134	30
Focusing	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 2: Counts of themes and intentions of messages in our corpus of 411 French texts; some texts carry more than one intentions and theme

focusing messages (named quantitative in [4]) 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 focusing 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 presentation (23%) but it is almost always possible to generate a focusing 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 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 distinguishes 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 did the canadian dollar fall?, 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 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 already appears in the graphic. 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 graphic 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.

3.1 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. Some types of messages can be generated using statistical computations: deductive messages using correlations, descriptive messages by giving the general trend of the curve of the data, domination messages by identifiying the extreme values, focusing messages by computing a mean or a sum of the data set but all of these do not solve the "how to say it problem".

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, like PostGraphe needs to know the intentions of the user, see 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 curves or to speak about the reversal of a tendency.

Data must also be identified with sufficient details 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 that information 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 [2]. 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 to 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 Text selection rules

After studying the corpus, we have identified text type selection rules for a writer's intentions. We also studied the comparison-evolution combination. All sentences given in this section were generated by SelTex, followed here by an English translation, done manually.

Comparison	
largest value	**if** its value is more than 10% higher than the second largest **and** there are more than 2 values
2 largest values	**if** the second highest value is more than 15% higher than the third largest **and** there are more than 4 values
3 largest values	**if** the third highest value is more than 20% higher than the fourth one **and** there are more than 6 values
smallest value	**if** its value is at least 30% lower than the second smallest one **and** there are more than 2 values
balance between values	**if** ratio between the largest and smallest value is less than 1.2 **and** there are more than 3 values
largest value	**if** none of the above

Table 3: Selection rules for a comparison

From the information shown in table 2, we can see that comparison is one of the most frequent use of information graphics, so we studied it quite extensively. Looking at the example data of Figure 2, we see that the first 3 highest values can be distinguished because there is a relatively large gap between the third (Ontario) and the fourth (Saskatchewan) and this is why SelTex generates `L'Alberta, la Colombie-Britannique et l'Ontario ont un taux plus élevé de ménages possédant un ordinateur` (*Alberta, British-Columbia and Ontario have a higher rate of households with a computer*). The last part, speaking of Québec appears because the user has indicated a particular interest in this value.

So for the comparison intention, we determined from our corpus study the empirical rules given in table 3[1]. The tests are done on the values of the data while excluding categories such as *others* and *total* that

[1]These rules are tried from top to bottom and SelTex outputs the first value on the left for which the conditions on the right are satisfied

are often found in information graphics.

Evolution	
increase / decrease	**if** all values are greater / smaller than the preceding one
most recent tendency	**if** there was a reversal of tendency
first and last values	**if** there were many reversals of tendency

Table 4: Selection rules for showing an evolution

Evolution occurs also quite often and for this, we use the rules given in Table 4 except when there are very big changes. For example, if the inflation rate in Canada from 1989 to 1995 were

$$\{5.0\%, 4.8\%, 5.6\%, 12.0\%, 1.8\%, 0.2\%, 2.1\%\}$$

then it would not be appropriate to compare the values of 1989 and 1995 because of the extremely high value in 1992. In this case, we only mention the final value such as in `Le taux d'inflation au Canada se situe à 2.1% en 1995` (*Inflation rate in Canada is at 2.1% in 1995*). In the case of a reversal of tendencies, we only pinpoint this fact; but if there were many reversals, we only give the starting and ending values. This is justified by the fact that the graphic shows the evolution over a long period of time while the text pinpoints what we think is of most interest to the reader: the last news.

Comparison and evolution	
evolution of the gap between curves	**if** there are two groups of serial data or the user is interested in two specific data
comparison of the last values of each curve	**if** there were many reversals of tendency
enumeration of the evolutions of each group	**if** otherwise

Table 5: Selection rules for showing the combination of comparison and evolution

We distinguish two kinds of combination:

evolution of comparison such as in `Le pourcentage de ménages possédant un ordinateur au Québec se rapproche de celui de l'Ontario` (*The rate of households with a computer in Québec is nearing the one of Ontario*)

comparison of evolution such as in `Le taux de ménages possédant un ordinateur à domicile augmente plus rapidement à Terre-Neuve qu'à l'Île du Prince-Édouard` (*The rate of households with a home computer increases faster in Newfoundland than in Prince Edward Island*)

Correlation

To determine if there is a statistically significant correlation between two sets of data, we use a standard linear correlation computation and then indicate if there is a positive or negative relation between the variables. But great care must be used in indicating positive or negative relations between variables so that they do not give false implications. For example, the following was encountered in our corpus `L'opinion des femmes quant à l'importance d'occuper un emploi rémunéré pour être heureuses varie`

`selon l'âge` (*The opinion of women about the importance of having a paid job to be happy varies according the the age*). Should we have written `augmente avec l'âge` (*increases with age*), we could have implied that women change opinion as they get older.

Distribution

In this case, we only implemented the case that indicates the interval where the values are highest.

Presentation

SelTex creates generic titles from the data description given by the user; if there are temporal data, then it also gives the starting and ending values such as in `Taux d'inflation au Canada de 1989 à 1995` (*Inflation rate in Canada from 1989 to 1995*).

4.1 Coverage of SelTex

We have implemented text schemas for 26 out of the 55 (47%) types of texts in our corpus but the ones we chose were the most frequent ones and they cover 71% of all texts in our corpus.

SelTex has been implemented in Prolog and is described further in [4]. The implementation is mainly based on the Prolog implementation of Frana [2] by Sylvie Giroux and Michel Boyer.

5 Related works

Our system is not the first one to combine text and graphics but the approach most closely related to our work is described by Green et al.[8] who describe a media-independent knowledge representation scheme for describing the content of communicative goals and actions. This language can represent information about focusing relations such as percentages, agregate and time dependent relations. It makes use of a first order logic with restricted quantification so it is much richer than our simple intention based language. On the other side, it requires more specific input from the user to create an appropriate graphic with the accompanying text to create a presentation.

Mittal et al.[10] describe a system to produce captions for complex charts. This system determines the content and structure of the captions by analysing the structure of graphical representations and the perceptual complexity of its elements and by using linguistic transformations such as ordering, agregation and centering. Although the generated captions are much more complex than the ones produced by SelTex, it seems better to produce simpler graphics and captions directly from the data and the user intentions rather than producing so complex graphics that they need an equally complex caption to be understood.

Other systems also use complementary texts and graphics. LFS [9] produces bilingual statistics on the unemployment and active population levels but the graphics and texts are almost always of the same predetermined type. Their syntactic realization component is better principled than ours though. FOG [7] generates bilingual graphics and text weather forecasts from numerical data. Another related system is COMMENT [1] which generates financial analysis reports with numerical data tables explained by written comments. The system uses a limited set of statistical rules to select a limited set of statistical rules to select a set of predefined comment types of 3 types: comparison with the mean, horizontal and vertical analysis. The results are very good but are quite specific for the application of financial analysis of townships.

6 Conclusion

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

[1] A. Bridault. Génération automatique de commentaires en langage naturel, 1995.

[2] C. Contant. Génération automatique de texte: application au sous-domaine du langage boursier français. Master's thesis, Département de linguistique et de philologie, Université de Montréal, 1985.

[3] M. Corio and G. Lapalme. Integrated generation of graphics and text: a corpus study. In M. T. Maybury J. Pustejovsky, editor, *Workshop on Content Visualization and Intermedia Representation (CVIR'98)*, pages 63–68, Montréal, August 1998. Coling-ACL'98.

[4] Marc Corio. Sélection de l'information pour la génération de texte associé à un graphique statistique. Master's thesis, Département d'informatique et de recherche opérationnelle, Université de Montréal, 1998.

[5] M. Fasciano and G. Lapalme. Intentions in the coordinated generation of graphics and text from tabular data. *soumis à Knowledge and Information Systems*, page 27p., August 1998.

[6] Massimo Fasciano. *Génération intégrée de textes et de graphiques statistiques*. PhD thesis, Université de Montréal, 1996.

[7] E. Goldberg, N. Driedger, and R. I. Kittredge. Using natural-language processing to produce weather forecasts. *IEEE - Expert*, 9(2):45–53, April 1994.

[8] N. Green, G. Carenini, S. Kerpedjiev, S. Roth, and J. Moore. A media-independent content language for integrated text and graphics greneration. In M. T. Maybury J. Pustejovsky, editor, *Content Visualisation and Intermedia Representations (CVIR'98)*, pages 69–75, Montréal, August 1998. Coling-ACL'98.

[9] L. Iordanskaja, M. Kim, R. Kittredge, B. Lavoie, and A. Polguère. Generation of extended bilingual statistical reports. *Proceedings of the* 15^{th} *International Conference on Computational Linguistics (COLING-92)*, pages 1019–1023, 1992.

[10] G. Carenini S. Roth V. O. Mittal, J. D. Moore. Describing complex charts in natural language: A caption generation system. *Computational Linguistics*, 24(3):431–467, September 1998.

[11] Gene Zelazny. *Dites-le avec des graphiques*. InterÉditions, 1989.