EnviroInfo 2007 (Warschau)
Environmental Informatics and Systems Research
 ISBN: 978-3-8322-6397-3

Text Planning of Air Quality Information

Nadjet Bouayad-Agha[1] , Leo Wanner[2]

Abstract

Two of the premises of the Multimodal Air Quality Information System known as MARQUIS are that (1) air quality information is dynamic (i.e., it might change every hour) and different data variations should be expressed differently, (2) people have different expertises, needs and interests about air quality and this must be reflected in the content and style in which this information is conveyed to them. The text planning module in MARQUIS achieves precisely that: it selects the content that is of relevance to a specific user from the mass of air quality information produced by the interpretation module, and arranges this information into a coherent discourse that takes into account the dynamic content and the profile of the user in the required platform. In this paper, we describe how we addressed these issues of dynamic content selection (section 2) and user modelling (section 3) in the text planning system of the MARQUIS project. These play an active role in the production of the ouput text plan (section 4) as they contrain the order of propositions and the types of discourse relations that occur between the propositions (section 5). We describe the architecture of the text planner (section 6) before giving some conclusions (section 7).

1. Introduction

Two of the premises of the Multimodal Air Quality Information System known as MARQUIS (Wanner et al, 2007) are that (1) air quality information is dynamic (i.e., it might change every hour) and different data variations should be expressed differently, (2) people have different expertises, needs and interests about air quality and this must be reflected in the content and style in which this information is conveyed to them. The text planning module in MARQUIS achieves precisely that: it selects the content that is of relevance to a specific user from the mass of air quality information produced by the interpretation module, and arranges this information into a coherent discourse that takes into account the dynamic content and the profile of the user in the required platform[3] (Bouayad-Agha et al, 2006).

Text planning is a fundamental step in Natural Language Generation (NLG) systems (Dale and Reiter, 2000). When the input data is time-series information describing weather forecast (Goldberg et al, 1994), stock market (Kukich, 1983) or gas turbine (Yu et al, 2006) activity, an important step is how to *summarise* this numerical data to form relevant messages. This requires expert knowledge of the domain. In addition, it has been long acknowledged in the field that the natural language output should be customised to the needs and profile of the target user (Zuckerman and Litman, 2001). For instance, a naïve user requires more detailed information and explanations than an expert user (Paris, 1993), and health information must be tailored to the patient's medical record and history and other personal characteristics (Hirst et al., 1997).

In this paper, we describe how we addressed these issues of dynamic content selection (section 2) and user modelling (section 3) in the text planning system of the MARQUIS project. These play an active role in the production of the ouput text plan (section 4) as they contrain the order of propositions and the types

[1] Department of Information and Communication Technologies, University Pompeu Fabra, Barcelona, Spain. e-mail: nadjet.bouayad@upf.edu
[2] ICREA and Department of Information and Communication Technologies, University Pompeu Fabra, Barcelona, Spain. e-mail: leo.wanner@upf.edu
[3] MARQUIS produces information tailored to different platforms (internet, email, newspaper, SMS), for five European regions (Catalonia in Spain, Finland, Upper Silesia in Poland, Baden Württemberg in Germany, Portugal) and in eight different languages (Catalan, Spanish, French, German, Polish, Finnish, Portuguese, English).

of discourse relations that occur between the propositions (section 5). We describe the architecture of the text planner (section 6) before concluding (section 7).

2. Content selection

In MARQUIS, "raw"' measured and forecasted pollutant concentration time-series and meteorological conditions time-series are delivered to the database of the service by regional monitoring networks. The pollutant index series and air quality index series are calculated and further interpreted by an air quality expert system shell, the "Interpretation and Assessment Module" (IAM) (Nicklaß, 2007). This module, designed by an air quality expert, produces the maximal air quality information that might be of relevance to any given user for any given region. This includes current, forecast and archive information, each of which contain pollutants' concentrations and air quality indices as well as more interpretive data such as important values (e.g., minima and maxima), relevant curve behaviour (e.g., drops, sharp changes) and exceedances. The text planner is in charge of selecting this content according to the user region and other profile characteristics, injecting region- and language-specific information and building the set of facts that will be included in the final text plan.
Region- and language-specific information includes the following:

1. **Alarm thresholds.** Mapping between a pollutant threshold value and label (e.g., information, alarm, 1 hour mean).
2. **AQ/Pollutant ratings.** Mapping between a pollutant concentration interval and a semantic label such as "good", "bad", "satisfactory", etc, and between an air quality index interval and a numerical scale such as "1/6" or "1/5".
3. **Health warnings.** Mapping between a pollutant concentration interval and a canned text health warning in all the languages of MARQUIS for each different type of user (e.g., doctor, general public).
4. **Time intervals.** Mapping between a time interval and a linguistic expression, such as "morning", "late morning", etc.
5. **Communicative significance.** Threshold of communicative significance of AQ index, that is, the value at which it becomes interesting to talk about an index (below which it is not worth mentioning).
6. **Delta levels.** Mapping between delta of AQ index or pollutant concentration and semantic labels such as "unchanged", "slightly" or "noticeably".

3. User modelling

A survey carried out within MARQUIS (Marquis-D3.1, 2005) revealed that a default user model that is based on the generalized personal data record of the user in question is adequate as a starting model. However, the user must be given the option to choose his default model and to individualise it. Such an individualisation is essential for the acceptance of the service by the user. Therefore, we foresaw a two-level user model. The first level consists of a hierarchy of default user profiles, which include at the first level domain professionals, medical specialists, patients, and the public. We have defined the default profiles for a number of user targets (e.g., general public, consultant and technician, doctors, asthmatic patients) in each of the different target platforms. These profiles describe the user needs, in particular, what AQ-information she/he wants, and for each one, the preferred modes. A user can change most of the settings of any of her/his profiles at any one time via a web interface. Example information to set in the user profile is whether the user requires details of the pollutants responsible for the air quality index, whether she/he wants detailed explanations, health warnings, what the temporal coverage of the archive information should be. Thus, we take advantage of both expert user modeling and the user's knowledge of

her/his own specific needs (Reiter et al, 2003). Figure 1 shows a simplified model of the default profile corresponding to the user type ``general public" and the platform ``web". This type of user requires for the web five segments of information: current information about Air quality (AQ) and pollutants (i.e., ime=today,threshold=false), alert information for AQ (i.e., time=today,threshold=true), forecast information for AQ (i.e., time=tomorrow,threshold=false) and archive information for pollutants (i.e., time=last week,threshold=false). A user can customize this model by deciding, for instance, that she/he is interested in archive information for Ozone only (greyed in the figure); she/he might also decide she is interested in archive information for the last month rather than the last week and that she/he prefers this information in a table and a text. Some settings however are set according to the type of user and cannot be modified by the user. For instance, for general public, the final text must provide a qualitative judgement on the air quality (i.e., judgment=true) but no quantitative value (i.e., value=false), and the pollutants information must be expressed with indices (i.e., value type=index in the profile element) if the home region of the user is Catalonia and concentration if her/his home region is Baden Württemberg. Thus, the idea is that all the information to be provided to a particular family of users is described explicitly in the user profile rather than hard-coded in the text planner.

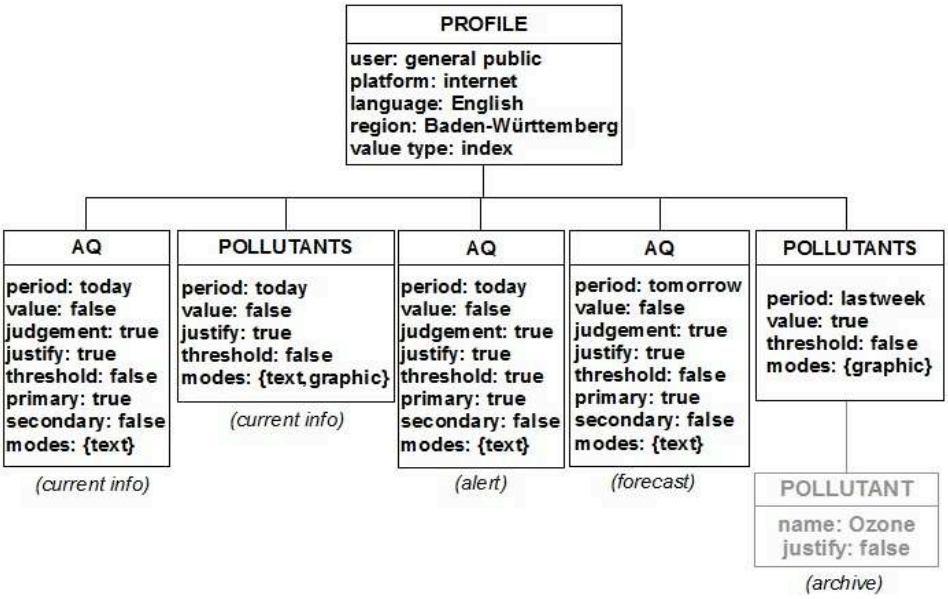
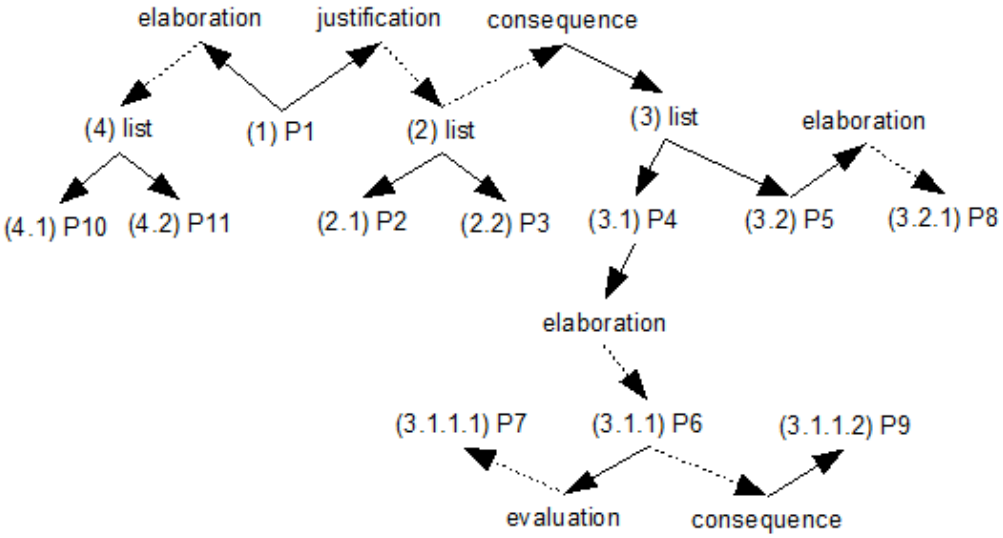


Figure 1: A simplified user profile for general public, web

4. Text plan

The facts selected according to the user profile are organised into sets of discourse tree fragments, each set forming a graph and related to a specific section of the final output (i.e., current information, alerts, archive, forecast). The nodes of the graph are either information units (the leaves of the graph) or spans, whilst the edges are either nucleus (the nucleus of a relation) or satellite (the supporting segment of a

relation), following Mann and Thompson (1988) Rhetorical Structure Theory (RST).4 Whilst RST structures are typically modellised into trees in NLG, several researchers; cf., e.g., (Bouayad-Agha, 2003) find that discourse graphs or mixed approaches combining trees and graphs (Webber, 2006) are more suitable for their domains and purposes, be it generation or analysis. Figure 2 shows an example of such a graph which is not equivalent to an RST-tree: contrary to "pure" RST, some propositions like P6 or the list of propositions P2+P3 are satellites to more than one relation. The navigation through the graph is ensured by the hierarchical order specified onto each information unit. This hierarchical order is not strictly isomorphic with the linear order, as the the partial repetition of P6 in the text shows ("Due to the ozone concentrations"), to circumvent the limitations of linearisation. The choice of discourse relations (e.g., JUSTIFICATION) and the ordering depends on the dynamic content and user profile, as described in the following section.



(P1) The air quality is bad because of **(P2+P3)** the lasting high temperatures and dry weather. (P4+P5) These make the ozone and PM10 levels rise. **(P6)** The ozone concentration reached this afternoon 190μg/m3, **(P7)** which is above the information threshold, and **(P8)** PM10 concentration 30μg/m3. **(P9)** Due to the ozone concentrations, people with weak heart conditions should avoid being outside for longer than absolutely necessary. **(P10+P11)** The low concentrations of the other prominent air pollutants (NO2 and SO2) do not influence significantly the air quality.

Figure 2: RST-like Discourse graph and corresponding text about air quality

[4] This text plan description applies to fragments of output that are realised as text. For graphics or tables, all the facts are encoded into a single information unit.

5. Constraints on text planning

Salience of numerical data and temporal information are one of the prime factors of dynamic content affecting discourse planning, in particular the relation of AQ and pollutant values with thresholds, comparison between so-called Very Important Points and distribution over a period of time (same day at different periods or consecutive days at same periods). Type of user (e.g., professional versus general public) but also user preferences are also determining factors of discourse planning. For instance:

- If the AQ index is above an alarm threshold, then health warnings must be provided first, while it is more appropriate to issue precautionary warnings after the elaboration on AQ.
- Low risk health warnings are presented in an ELABORATION relation with AQ index whilst high risk health warnings are presented in a CAUSE relation: "*Today, air quality is fair. There are no harmful effects to human health expected"* vs *"Increase of reversible short term effects to human health is likely with sensitive people. This is caused by today's bad air quality"*.
- If the delta between minimum and maximum is significant, then the relation is CONTRAST; if it is unchanged, then it is ANALOGY; otherwise it is LIST.
- If the AQ index is suboptimal, then its relation with primary pollutants is CAUSE; otherwise it is ELABORATION.
- The relation between AQ rating and secondary pollutants (i.e., pollutants that do not contribute to the AQ) is ELABORATION if the user is a professional (e.g.,*"The air quality is bad. NO2 does not contribute to it")*, and CONCESSION if the user is general public (e.g., *"The air quality is bad. However, NO2 does not contribute to it"*).

6 Discourse The text planner at work

Figure 3 shows the architecture of the text planner, which is divided into the traditional pipeline content selection and discourse structuring modules, each of which are further divided into a set of tasks. Each task is implemented as a specialised set of rules, or templates, to use the XSLT terminology. XSLT is a powerful language that can be used for transforming one or more XML inputs into an XML output, and therefore it was particularly suitable for the task at hand. Indeed the user profile, assessment and other system's input are all in XML format to facilitate intercommunication between modules (and project teams). The sets of templates associated to each task are then called in a pipeline with their corresponding inputs and outputs. Once the data representations are agreed upon, this architecture allows for rapid and iterative development and updating of the system. This technology has already been used successfully for NLG in general (Wilcock, 2001) and text planning in particular (Foster, 2004). In what follows, we describe in more details the main steps of the text planner.

6.1 Facts production

This task is divided into three separate steps which are (1) region localisation, (2) simple facts production and (3) complex facts production. Region localisation chooses those portions of the assessment data which matches the user's region specified in the profile. Indeed most of the assessment data is region-specific, such as the air quality or pollutant indices. Simple and complex facts production are in charge of gathering the dynamic and static input data together in *facts* which will then be used as information units (i.e., leaves) in the discourse structure. The complex facts production step uses the facts produced in the simple facts production step to produce more elaborate facts, for instance, the set of facts on primary pollutants based on the set of facts on all pollutants. The advantage of this cascaded approach is that the related entities will corefer, such as the pollutant alarm threshold and its concentrations which have the

same value, or a primary pollutant in the list of primaries and a pollutant in the list of pollutants. This coreference of entities will be used by the generators to produce textual references between spans.

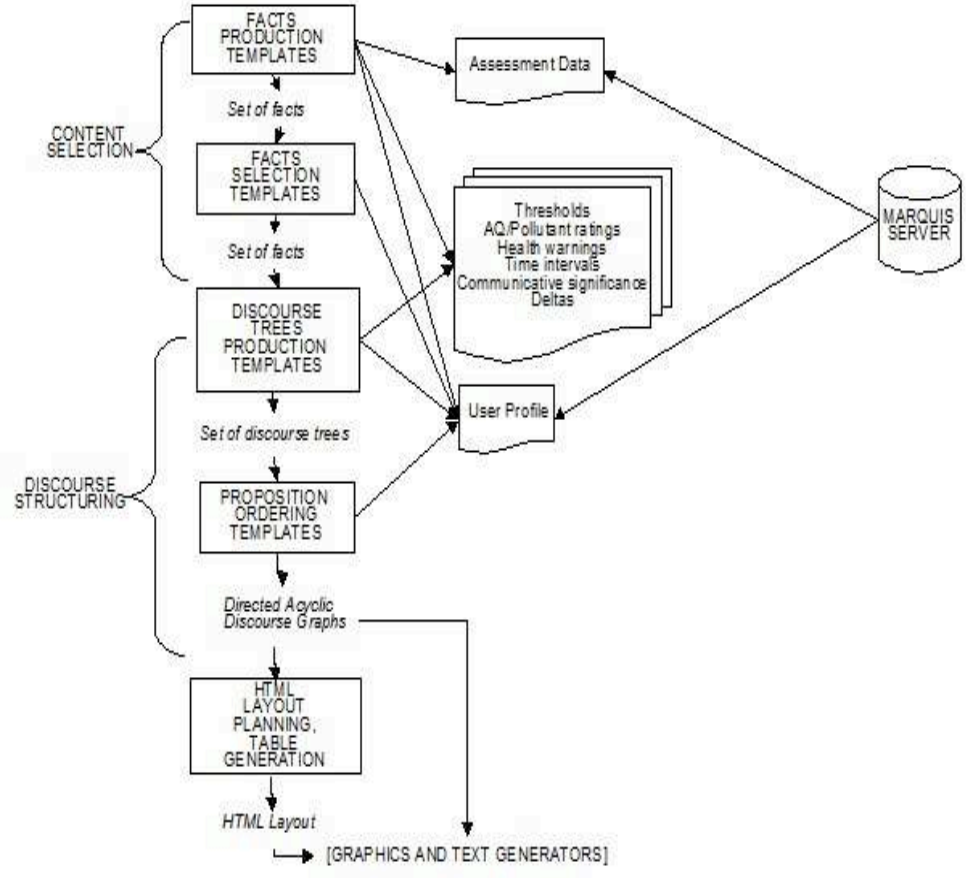


Figure 3: Architecture of the Text Planner

6.2 Facts selection

The facts selection task selects the facts according to the user profile. It is divided into four sets of templates, each responsible for the selection of a specific segment of information, namely, current information, alerts, forecast and archive. The output of this task are four separate xml files, which will continue to be processed separately in subsequents stage and will be merged into a single xml output only at the end. In addition to selecting the content according to the user profile, this task is in charge of informing the user if information is missing, as shown in the following gloss of the alert template:

```
IF  user wants alert info THEN
     IF assessment has daily info THEN
       IF there is threshold info THEN
        include the threshold info
       ELSE
        include <nodata ref="no_alerts"/>
       ENDIF
     ENDIF
   ELSE
     include <nodata ref="miss_asst"/>
   ENDIF
```

6.3 Discourse trees production

Each section (current information, alerts, archive, forecast) calls a set of templates that specify the possible mappings between semantic and discourse relations that apply to a pair of facts according to dynamic content and user profile constraints such as the ones described in section 5. In some cases these constraints impose a specific order on the nucleus and the satellite by virtue of the nature of the relation or otherwise more forcefully on a relation that normally allows either order. This is illustrated by the relation between a concentration and an alert threshold which can be EVALUATION, in which case the concentration must be presented before the threshold (*"The Ozone concentration reached 190 µg/m3 this afternoon. This means that the information threshold was exceeded.");* or EVIDENCE in which case either order is possible (e.g., *"The Ozone information threshold has been exceeded. Indeed, the Ozone concentration reached 190 µg/m3 this afternoon.")*. In this latter case, when an alarm threshold is reached, then it must be presented before the concentration.

6.4 Propositions ordering

Once the discourse tree fragments have been produced together with some fixed or flexible ordering between the propositions within each tree, the fragments must be ordered with respect to one another. For instance, the tree fragment threshold-health-risks must appear before the tree fragment concentration-threshold if the threshold is "alarm". If the user has expressed a special interest for a specific pollutant (i.e., s/he is a asthmatic or allergic to that pollutant) in addition to requesting information about the air quality's primary and secondary pollutants, then information about that specific pollutant must be presented first.

7. Conclusions

Our approach to text planning allows us to take into account dynamic content and individual user requirements at every stage of the planning process, from the selection of facts, to the choice of discourse relations and the ordering. This, we hope, makes the communication of air quality information more effective to the end users.

Bibliography

Bouayad-Agha, N. (2003): Non-hierarchical Planning of Document Structures. Proceedings of EACL'03. Budapest, Hungary.

Bouayad-Agha, N., Wanner, L., Nicklaß, D.(2006): Discourse Structuring of Dynamic Content. In: Proceedings of Sociedad Española de Procesamiento de Lenguaje Natural - SEPLN'2006, Zarragoza.

Foster, M.H., White, H. (2004): Techniques with Text Planning with XSLT. Proceedings of NLPXML'04.

Goldberg, E., Driedger, N., Kittredge, R. (1994): Using Natural-Language Processing to Produce Weather Forecasts. IEEE Expert: Intelligent Systems and Their Applications. 9:2 pages 45-53.

Hirst, G., DiMarco, C., Hovy, E.H., Parsons, K. (1997): Authoring and Generating Health-Education Documents That are Tailored to the Needs of the Individual Patient. Proceedings of the 6th International Conference on User Modeling.

Kukich, K. (1983): Knowledge-Based Report Generation: a technique for automatically generating natural language reports from databases. Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval.

Mann, W.C, Thompson, S.A. (1988): Rhetorical Structure Theory: A theory of text organization. in L. Polanyi (ed.) *The Structure of Discourse*, Ablex.

Marquis-D3.1 (2005): Specification of the User Profiles Report. Marquis Project Deliverable 3.1.

Nicklaß D., Bouayad-Agha N., Wanner L. (2007): Addressee tailored interpretation of air quality. In: Proceedings of EnviroInfo 2007, Warsaw.

Paris, C. (1993): User Modelling in Text Generation. Frances Pinter Publishers.

Reiter, E., Dale, R. (2000): Building Natural Language Systems. Cambridge University Press. Reissued in paperback in 2006.

Reiter, E., Sripada, S., William, S. (2003): Acquiring and Using Limited User Models in NLG. Proceedings of 2003 European Natural Language Generation Workshop, pp. 87-94.

Wanner, L., Nicklaß, D., Bohnet, B., Bouayad-Agha, N., Bronder, J., Ferreira, F., Friedrich, R., Karppinen, A., Lareau, F., Lohmeyer, A., Panighi, A., Parisio, S., Scheu-Hachtel, H., Serpa, J. (2007): From Measurement Data to Environmental Information - MARQUIS - A Multimodal AiR QUality Information Service for the General Public. In: Swayne A., Hrebicek J. (Eds.): Environmental Software Systems – Dimensions of Environmental Informatics, Vol.7.

Webber, B. (2006): Accounting for Discourse Relations: Constituency and Dependency. In M. Butt, M. Dalrymple and T. King, Intelligent Linguistic Architectures, Stanford: CSLI Publications,pp. 339-360.

Wilcock, G. (2001): Pipelines, Templates and Transformations: XML for Natural Language Generation. Proceedings of NLPXML'01

Yu, J., Reiter, E., Hunter, J., and Mellish, C. (2006): Choosing the content of textual summaries of large time-series data sets. Natural Language Engineering, 13: 25-49

Zukerman, I., Litman, D. (2001): Natural language processing and user modeling: Synergies and limitations. User Modeling and User-Adapted Interaction, 11: 129-158.