Discourse Structuring of Dynamic Content*

Nadjet Bouayad-Agha
Technology Department
University Pompeu Fabra
Barcelona
nadjet.bouayad@upf.edu

Leo Wanner
ICREA and Tecnology Department
University Pompeu Fabra
Barcelona
leo.wanner@icrea.es

Daniel Nicklass
IER
University of Stuttgart
Stuttgart, Germany
dn@ier.uni-stuttgart.de

Resumen: Uno de los desafíos de la Generación de Lenguaje Natural es la adaptación de la estructura y las palabras de la salida lingüística a la habilidad del usuario, el contenido, el género apropiado, el estilo, etc. Nos centramos en la determinación de la estructura del discurso. En general, se supone que entre dos unidades de contenido ocurre siempre la misma relación de discurso. Propuestas que varían el tipo de relación discursiva y el orden de las proposiciones según la interpretación del contenido siguen siendo escasas. Sin embargo, tal interpretación es extremadamente importante especialmente si el contenido es altamente dinámico como por ejemplo, cuando los datos son series temporales. Presentamos un planificador de textos que considera las restricciones que imponen los datos dinámicos para tomar decisiones a cada etapa de la planificación, en particular para la selección de las relaciones discursivas y la ordenación de las proposiciones.
Palabras clave: generación de textos, contenido dinámico, planificación RST, XSLT, información sobre la calidad del aire

Abstract: One of Natural Language Generation's continuing challenges is to determine the structure and words of the generated linguistic output in accordance with the expertise of the user, the content, the appropriate genre, style, etc. We focus on the determination of the discourse structure. Most often, it is assumed that between two content units always the same discourse relation holds. Approaches in which the choice of discourse relations and the ordering of propositions depends on the interpretation of the content are still scarce. However, such an interpretation is extremely important especially if the content is highly dynamic as, e.g., in the case of data parameter time series. We present a text planner that takes into account the constraints imposed by dynamic data to make decisions at every stage of the text planning, and in particular, for the selection of discourse relations and the ordering of propositions.
Keywords: text generation, dynamic content, text planning, RST, XSLT, air quality information

1 Introduction

One of Natural Language Generation's continuing challenges is to adapt the form and wording of the generated linguistic constructions to the expertise of the user, the content, the appropriate genre, style, etc. In text planning, content selection and discourse structuring are tightly intertwined (Dale and Reiter, 2000). User models (Paris, 1993) and discourse history for hypertext NLG (O'Donnell et al, 2001) give some flexibility to the text planner, triggering the selection of different rhetorical schemas or rhetorical assertions between facts. Approaches that aim at introducing flexibility in the derivation of the structure of the discourse (rather than that of the content) are still scarce. Kosseim and Lapalme (2000), for instance, investigate a many-to-many mapping between semantic and rhetorical relations and provide some general constraints, while Vander Linden and Martin (1995) propose a network of intrinsic semantic constraints for ordering as well as realising the "purpose" relation.

Our work in the domain of the generation of multimodal, multilingual air quality information has shown that if the content is of a highly dynamic nature (it may change each time the

* The work reported on in this paper has been carried out in the framework of the MARQUIS-project funded by the European Commission in the framework of the eContent programme under the contract number EDC-11258; duration: 2005-2007. We would like to thank all colleagues of the MARQUIS-Consortium, and in particular Bernd Bohnet, for their valuable help.

concentration of the individual air pollutants is measured—e.g., every hour), the ordering of the propositions and the discourse relations between the propositions largely depend on its interpretation. For instance, if proposition P_1 states the highest concentration of an air pollutant at a given time and proposition P_2 states the rating (such as 'excellent', 'good', ..., 'very bad') of air quality, which is directly related to this concentraion, then the discourse relation between P_1 and P_2 (and their ordering) depends on the assessment of this rating. Consider an example:

1. LIST: [*The primary pollutant today is ozone with 30$\mu g/m^3$.*]$_{P_1}$ [*The air quality is good.*]$_{P_2}$
2. CAUSE: *Due to the high ozone concentration of more than 180$\mu g/m^3$.*]$_{P_1}$, [*the air quality is bad*]$_{P_2}$

That is, if the air pollution is low, it is considered more adequate to present the statement on air quality (i.e., P_2) either as a further statement that follows the statement on ozone concentration (leaving the causal relation between the ozone concentration and the overall air quality implicit) or as a consequence of the ozone concentration. In the first case, the discourse relation between $_{P_1}$ and $_{P_2}$ is LIST (as shown above); in the second case, it is CONSEQUENCE:

1.′ CONSEQUENCE: [*The primary pollutant today is ozone with 30$\mu g/m^3$.*]$_{P_1}$ [*The air quality is <u>thus</u> good.*]$_{P_2}$

If the air pollution is high, it is considered adequate to make explicit the causal relation between the ozone concentration ($_{P_1}$) and the air quality ($_{P_2}$).

In the case of sufficiently bad air quality conditions, it is even more appropriate to reverse the order of $_{P_1}$ and $_{P_2}$—which let appear $_{P_2}$ in the focus:

2′ JUSTIFICATION: [*Currently, we experience a bad air quality episode,*]$_{P_2}$ [*which is due to high ozone concentrations of more than 200$\mu g/m^3$.*]$_{P_1}$

As the example shows, the change of the order of propositions usually also implies a change of the discourse relation.

In what follows, we present a text planner that takes into account the constraints imposed by the dynamic data to make decisions at every stage in the text planning, including choice of discourse relations and ordering of the propositions.

The remainder of the paper is structured as follows. In Section 2, present first the general setting of our work and focus then on the interpretation module that is in charge of transforming the raw data into interpreted data ready to be used by the text planner. In Section 3, we discuss the constraints that the dynamic content puts on discourse planning, before describing the text planner at work (Section 4), and giving our conclusions and outlining some future work for the project and beyond (Section 5).

2 The Framework

The text planner we report here on is a module of the *Multimodal AiR Quality Information Service for General Public* (MARQUIS) for the automatic delivery of multilingual periodic information about air quality via different communication platforms: web, mobile services (SMS, WAP, MMS), email, television and printed media. As modi, text, tables, and graphics are used. MARQUIS covers five different European regions (Baden-Württemberg in Germany, Catalonia in Spain, Finland, Portugal, and Upper Silesia in Poland). The information is being produced in eight different languages (in addition to the first language of each region, we generate English, French and Spanish) and takes regional characteristics, typological features of different user profiles, and style restrictions of each communication platform into account.

2.1 General Setting: MARQUIS

MARQUIS realizes a "two-pipe" architecture. In the first pipe, the air quality data (concentrations of main air pollutant substances such as O_3 (ozone), PM_{10} (dust particles), NO_2 (nitrogen dioxide) and CO (carbon monoxide) measured at the monitoring sites distributed over the above regions are periodically received at the MARQUIS-server and assessed and interpreted by an assessment module. In the second pipe, the selection of the content relevant to a specific user, discourse planning and information generation takes place. To start content selection, the text planner is triggered by user information requests. Once a request is detected by the MARQUIS-user interface, the text planner receives from the server the profile of the user in question and the assessment data for the required site. The text planner produces a text plan, assigning to the sentence table and graphics generators the text plan fragments to be realized in the correspond-

ing mode. After the chunks of information are generated by the corresponding generators, they are merged together into a single document and delivered to the user via the MARQUIS-client interface.

The role of the text planner is thus to take care of all the linguistic information associated with the air quality data (e.g., associating canned text to alarm thresholds) and to provide the generators with *all and only* the information—conceptual, discourse and otherwise—they need for producing the requested documents.

2.2 The Interpretation Module

The introduction of the interpretation module is important for the understanding of the text planning strategy in MARQUIS. Like in many generation systems summarizing time-series data (Yu et al, 2006), the interpretation module is in charge of assessing the raw data, making the necessary interpretations and extracting relevant information that may be communicated to the user and presenting it in a format usable by the text planner.

The two major types of raw data are the observed and forecasted meteorological parameters (such as precipitation, wind strength, wind direction,temperature, etc.) and the observed and forecasted concentrations of the main pollutants (O_3, PM_{10}, NO_2 and CO) of all monitoring networks in terms of value-time series.[1] The value-time series are further enriched by contextual characteristics (season, day of the week, time of the day, type of the monitoring site, etc.) that are of relevance for the assessment.

The assessment output structure produced by the interpretation module contains the maximum content that can be derived from the raw data available at a given time. The structure is divided into three main sections: ARCHIVE (archival data for any number of days), DAY (current day's data—partly observed and partly forecasted) and FORECAST (data for the next day). Each of the main sections can include an element called POLLUTANT. Within this element, the rating of a pollutant's concentration on a quality scale and the meteorological explanation for this rating (on the given day) are stored. Embedded elements contain the following information:

Time concentration tuples. The distribution of the concentrations of the pollutants over a given day (henceforth, "course-of-day").

Exceedence sequence. How often and how long a threshold was/will be exceeded, what type of threshold it is (i.e., information or alarm) and what was/will be the (absolute and relative) maximal exceedance.

VIP sequence. The most prominent points in the course-of-day curve such as first, current, minimum, maximum and average values of a day.[2]

Delta sequence. The concentration value differences between specific point pairs (such as minimum and maximum, first and current, etc.). These differences are labelled with a qualitative interpretation: *unchanged, slight, strong, etc*.

VIC sequence. The most prominent changes in the course-of-day, which are specified by three attributes: GRADIENT (*sharp* or *slight*), TENDENCY (*raise* or *drop*), and STEADINESS (*true* or *false*).[3]

For the ARCHIVE and DAY sections, two additional elements are available, namely "AQ", which covers the air quality indices for all regions, and "UV", which includes the minimum and maximum ultraviolet index.

The assessment output structures (as the user profile specifications and all other static information used by the system) are in XML-format to facilitate intercommunication between modules (and project teams).

3 Dynamic Content-based Constraints

Salience of numerical data and temporal information are one of the prime factors for dynamic content affecting discourse planning, in particular the relation of AQ and pollutant values with thresholds, comparison between VIPs and distribution over a period of time (same day at different periods or consecutive days at same periods). For instance:

If the AQ-index is above an alarm threshold, then health warnings are provided first; while it is more appropriate to issue precautionary warnings after the elaboration on AQ. In addition, low risk health warnings are presented in an ELABORATION relation with AQ index while high risk health warnings are presented in a CAUSE relation: *Today, air quality is fair. There are no harm-*

[1] The frequency with which the concentrations are measured varies from hourly to daily.

[2] VIP stands for "Very Important Point".

[3] VIC stands for "Very Important Change".

ful 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 difference 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.

4 The Text Planner at Work

As Figure 1 shows, the text planner is divided into the traditional pipeline content selection and discourse structuring modules, each of which is further divided into a set of tasks. Each task is implemented in XSLT as a specialised set of rules (or *templates*). 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. The sets of templates associated to each task are then called in a pipeline with their corresponding inputs and outputs using ANT[4], a platform-independent XML-based scripting language. 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, 2003) and text planning in particular (Foster and White, 2004).

The figure shows that, in addition to the assessment (or dynamic) data, there is a number of static inputs which are used as input to the text planner. They are concerned with providing some form of region-specific linguistic interpretations to the assessment data (whether through canned text or semantic labels). Most of them are required and provided by the European Commission and national authorities.

Concentration-to-Index mappings. Assigning concentration ranges to AQ-index and pollutant-index scales. This is useful when AQ-information is given in terms of indices rather than concentrations.

Index ratings. The correspondance between intervals of AQ and pollutants values and ratings such as "good", "bad", "satisfactory", etc.

Delta levels. The correspondance between intervals of AQ and pollutants differences and labels such as "unchanged", "slight", "strong", etc.

Alarm thresholds. For each pollutant, give a canned text message for a given threshold.

Health warnings. For each pollutant, give a canned text health warning for each concentration level for each different type of user in all different languages.

Time intervals. Map a time interval to a linguistic expression, such as "morning", "late morning", etc.

Meteo justification. Provide a set of meteorological justifications ("dry weather", "high temperatures", etc.) for the measured concentration of each pollutant.

Communicative significance. For AQ-index and each pollutant, give a threshold value which is communicatively significant (though below any alarm threshold).

The final output of the discourse structuring is a directed acyclic graph where the edges are rhetorical relations or propositions, as shown in Figure 2 for a target text. The hierarchical order is indicated on the propositions or the list relations (which are considered a conjunction of propositions). There are several motivations for building a graph rather than a tree—as is more traditional in text planning. Firstly, the sentence generator decomposes the text plan into a conceptual graph as it is based on the Meaning-Text Theory (Mel'čuk, 1988) so a tree would be superfluous. Secondly, a tree would require the use of artificial relations between disconnected parts of the discourse (e.g., different sections of the document) such as *joint*. As a matter of fact, after the facts production phase, the different sections (i.e., current info, forecast, archive, alert) are generated independently of one another. Finally, the rhetorical graph is less restrictive than a tree, since it does not rely on a nuclearity principle for its interpretation (Marcu, 1996; Bouayad-Agha, 2003). This is illustrated in the graph in the following places: P2+P3 act as satellite of both P1 and P4+P5, and P6 is satellite of both P4 and P9.

The tasks performed by the text planner are described below. The discourse relation specification is performed in the "Discourse Trees Production Task" and the propositions ordering in the "Propositions Ordering Tasks".

[4]Apache Ant 1.6.5 http://ant.apache.org/

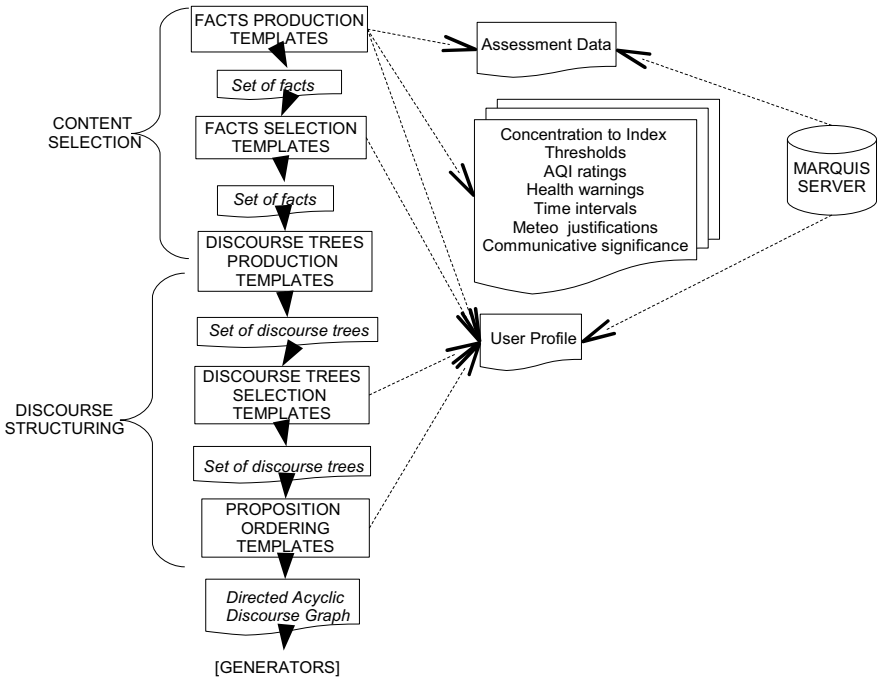

Figure 1: The Text Planner Architecture

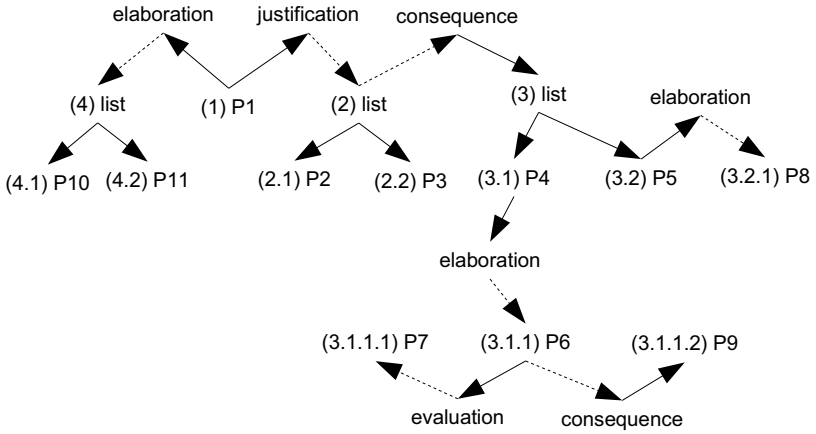

Figure 2: The Discourse graph for: *(P1)* The air quality is bad because of *(P2+P3)* the lasting high temperatures and dry weather. *(P4+P5)* These make the ozone and PM10 levels rise. *(P6)*The ozone concentration reached this afternoon 190g/m3, *(P7)* which is above the information threshold, and *(P8)* PM10 concentration 30g/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.

Fact Production Task

This task is divided into three separate steps which are: (1) region localisation, (2) simple fact production, and (3) complex fact production. Region localisation chooses those portions of the assessment data which match 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 fact productions are in charge of gathering the dynamic and static input data together in *facts*, to be be used as information units (i.e., leaves) in the discourse structure. The complex fact production step uses the facts produced in the simple fact production step to generate more elaborated facts—e,.g., 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; for instance, 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 is used by the sentence generator to produce textual references between spans.

Fact Selection Task

The fact 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 is subsequently processed separately in subsequents stage and merged at the end into a single XML-output. 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
```

Discourse Tree Production Task

For a given set of facts, all possible discourse trees are produced in all possible modes. For instance, the pollutant concentrations might be presented in a table *and* in a graphic, each with a different discourse tree. The choice of discourse relations according to rules like the ones described in Section 3 is performed at this stage. These rules are incorporated into the templates. Cf. the gloss of the template in charge of producing the relation between a pollutant and its health warning:

```
IF  pollutant rating is suboptimal THEN
  relation between pollutant
  concentration/index and health
  warning is ``cause"
ELSE
  relation between pollutant
  concentration/index and health
  warning is ``elaboration"
ENDIF
```

Discourse Tree Selection Task

This task's role is to select the discourse spans according to the preferred modes (graphic, table, text) specified in the user profile.

Proposition Ordering Task

This task's objective is to give the propositions in the discourse structure a hierarchical order. A gloss of an ordering template is:

```
IF  alarm threshold THEN
  express the health warnings first.
ELSE
  express the AQ information first.
ENDIF
```

For the discourse structure in Figure 2, starting from the proposition P_1, the hierarchical order described below is obtained. Note that the hierarchical order and the linear order are not strictly identical, since in the surface text, P_9 is extraposed with respect to P_6. In this latter case, extraposition is indicated in the text with the repetition *Due to ozone concentrations*. The necessity of such an extraposition results from information structure restrictions (*thematic progression*) dealt with apart. We will not discuss this issue further, although this point raises some interesting possibilities (e.g., an extraposition operation could be performed by disjoining the consequence(n:P_9,s:P_6) assertion and joining it in an ELABORATION relation with the LIST of primary pollutants P_4+P_5).

```
(1) AQI (P1)
(2) Correlating Meteo for primaries
  (2.1) High temperatures (P2)
  (2.2) Dry weather (P3)
(3) Primary pollutants
  (3.1) Primary Ozone (P4)
    (3.1.1) Ozone concentration (P6)
       (3.1.1.1) Info threshold (P7)
       (3.1.1.2) Health warning (P9)
  (3.2) Primary PM10 (P5)
    (3.2.1) PM10 concentration (P8)
(4) Secondary pollutants
  (4.1) Secondary NO2 (P10)
  (4.2) Secondary SO2 (P11)
```

5 *Conclusions and Future work*

Our approach to text planning allows us to drive the choice of discourse relations and ordering of propositions using external factors such as constraints resulting from the highly dynamic content. We have a powerful interpretation module that is able to periodically produce a set of data relevant for communication to the user. Our text planner architecture is divided into a set of well-defined tasks together with the use of a discourse graph. It is suitable for introducing constraints at specific points in the text planning, such as discourse relation determination and proposition ordering. Currently, the output text still lacks the fluidity of the target text shown in Figure 2, because sentence aggregation and coreference are still not implemented by the sentence planner, and the linear order is obtained simply by following the hierarchical order.

The work plan for the time to come foresees communicating more complex information to the user, such as comparative information between air quality in neighbour locations, today's and yesterday's, and tomorrow's and today's air quality at a given location, discussion of the evolution of air quality indices or pollutant concentrations throughout the day, etc. We are also planning to use our approach to text planning in other related domains—in particular meteorology (e.g.,maritime forecast) and to evaluate the different orders and rhetorical relations on target users in order to validate our rules.

References

Nadjet Bouayad-Agha. 2003. *Non-hierarchical Planning of Document Structures*. Proceedings of EACL'03. Budapest, Hungary.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Mary Ellen Foster and Michael White. 2004. *Techniques with Text Planning with XSLT*. Proceedings of NLPXML'04.

Leila Kosseim and Guy Lapalme. 2000. *Choosing Rhetorical Structures to Plan Instructional Texts*. Computational Intelligence: An International Journal. 16(3): 408–445. Blackwell. Boston.

Michael O'Donnell, Chris Mellish, Jon Oberlander and Alistair. 2001. *ILEX: an architecture for a dynamic hypertext generation system*. Natural Language Engineering 7:225–250.

Daniel Marcu. 1996. *Building up rhetorical structure trees.* Proceedings of AAAI-96. American Association for Artificial Intelligence.

Igor Mel'čuk. 1988. *Dependency Syntax.* SUNY Press.

Cécile Paris. 1993. *User Modelling in Text Generation*. Frances Pinter Publishers.

Keith Vander Linden; James H. Martin. 1995. *Expressing Rhetorical Relations in Instructional Text: A Case Study of the Purpose Relation*. Computational Linguistics, 21(1).

Graham Wilcock. 2001. *Pipelines, Templates and Transformations: XML for Natural Language Generation.* Proceedings of NLPXML'01.

J Yu, Ehud Reiter, J Hunter and Chris Mellish. 2006. *Choosing the content of textual summaries of large time-series data sets* To appear in Natural Language Engineering.