Generating Air Quality Reports From Environmental Data

Stephan Busemann, Helmut Horacek*
German Research Center for Artificial Intelligence (DFKI GmbH)
Stuhlsatzenhausweg 3, 66123 Saarbrücken (Germany)
{busemann, horacek}@dfki.de

Abstract

This paper describes ongoing work on the generation of German and French air quality reports on the basis of up-to-date environmental measurements. This real-world application is characterized by a simple and small sublanguage. The system is called with a bundle of user requests entered through a hyper-link navigator. For text planning, a schema-based component produces domain-specific semantic content representations that are fed to the TG/2 production system [Busemann, 1996] for linguistic realization. The semantics interface between the two components is tailored to the task and domain at hand. It is independent from the particular language chosen. It is argued that these design decisions have important practical benefits over more general, linguistic approaches. The texts produced are designed for administrative use. A version for the general public is foreseen as well.

1 The application scenario

This paper describes ongoing work on the generation of German and French air quality reports on the basis of regularly updated environmental measurements. Such data is made available on a server under development for TEMSIS (Transnational Environmental Management Support and Information System). It includes the pollutant, the measurements, the location and the time the measurements were taken, and thresholds that may cause some activity if overstepped. Besides such data, the server provides meta data that allow for descriptions of the measuring locations, of the pollutants measured and of regulations or laws according to which a comparison between measurements and thresholds can be performed.

With TEMSIS, an environmental information system is designed and implemented as part of a transnational cooperation between the communities in the German-French urban agglomeration, Moselle Est and Stadtverband Saarbrücken. Networked information kiosks will be installed in a number of communities to provide public and expert environmental information.

The timely availability of relevant information about the current environmental situation improves the planning and reactive capabilities of the administration considerably. The summarization of information in natural languages saves time by reducing the need of looking up heterogeneous data on the server. The domain of air quality reports is especially promising in this respect since the underlying data are relatively complex and highly structured. The generated texts can be complemented with graphical presentations of the development of measurements over time or comparisons with earlier periods. The generated texts can be edited by the administration to fit additional needs.

The generation system has interfaces to the server (access to data and meta data) and to the navigator, through which the user selects his request from a hierarchy of options.

*This work was funded by the European Union within the TEMSIS project (Telematics Applications C9, no. 2945).

Moreover, an interface to a diagram graphics generator is foreseen. The results are presented on the Web as an HTML document. The generation system must fulfill two tasks:

Text structuring: A schema-based component generates the text structure on the basis of the user's request. It combines fixed text blocks with dynamic text in an as language-neutral way as possible.

Surface realization: The production system TG/2 [Busemann, 1996] is reused for language-specific processing. It will operate with grammars partly tuned towards the domain and task requirements.

In addition, in order to be cooperative and helpful, the system must exhibit robustness in all its parts. Any of the interfaces in the generator may provide unexpected input. For instance, the number of available measurements may be insufficient to fit a schema, or the user request may be ill-formed. Nevertheless the generator has to capture such failures and produce adequate meta-level comments.

2 Text structuring using real world data

The overall task of the text planning component in TEMSIS is the production of an intermediate structure of an air pollution report suiting a small set of user specifications. These specifications determine a report structure and access paths to the concrete data to be included in the report. The report structure is taken out of a small set of pre-defined structures, which were defined on the basis of analyses carried out by domain experts in Germany and France. Each report consists of a set of assertions whose composition is obtained in varying degrees of cannedness (numbers refer to the sample report in Figure 1):

- Canned texts taken from the database (2), (6); these assertions constitute descriptions of the major domain concepts involved. Their inclusion into the report is optional.
- Freely generated, but data-independent assertions (1), (3); these assertions represent confirmations of user parameters. Their inclusion into the report is optional, too.
- Freely generated, data-dependent assertions (4), (5), (7); these assertions constitute presentations of stored or derived data, selected from the database in accordance with the user's specifications. Moreover, even the structure of these assertions depends to some extent on the database content; in case the data stored are considered insufficient for a reliable statement about the requested data, a suitable qualifying statement about the requested information is added. This is one of the places where cooperativeness meets robustness.

In addition, vectors of data are depicted as diagrams in some report types. Diagrams will not be produced by our system, but inserted into the text if appropriate at locations defined through the text structure. If a graphics component is not available, the information to be diagrammed can alternatively be presented as a formatted table.

In more technical terms, major tasks in the text organization part of the TEMSIS system are the following:

(1) Zur Betrachtung der Luftbelastung im Winter 1996/97 haben Sie die Meßstation Völklingen-City ausgewählt. *(optional)*

[In order to inform yourself about the air pollution during winter 1996/97, you have chosen the measurement station of Völklingen-City.]

(2) Die Lage der Meßstation Völklingen-City kann wie folgt charakterisiert werden: Die Station liegt mitten in der Völklinger Innenstadt auf 220 Meter Meereshöhe. Gemessen wird in 4 m Höhe über dem Boden. Die Station ist von Gebäuden umgeben und liegt an einer stark befahrenen Straße. *(optional)*

[The location of the measurement station of Völklingen-City can be described as follows: It is located ... at 220m of altitude...]

(3) Sie wollen sich über die Konzentration von Schwefeldioxyd in der Luft informieren. *(optional)*

[You want to know the concentration of sulfur dioxide.]

(4) Im Winter 1996/97 wurde der MIK-Wert nach VDI-Richtlinie 2310 von 1000 $\mu g/m^3$ an der Meßstation Völklingen-City nicht erreicht. Der MIK-Wert für eine 24-stündige Einwirkungsdauer (300 $\mu g/m^3$) wurde dreimal überschritten.

[During the winter 1996/97, the MIK value according to VDI directive 2310 of 1000 $\mu g/m^3$ was not reached at the measurement station of Völklingen-City. The MIK value for an exposition of 24 hours (300 $\mu g/m^3$) was exceeded three times.]

(5) Im Winter 1995 wurde der MIK-Wert nicht erreicht. Der MIK-Wert für eine 24-stündige Einwirkungsdauer wurde einmal überschritten.

[During winter 1995/96, the MIK value was not reached. The MIK value for an exposition of 24 hours was exceeded once.]

(6) Schwefeldioxyd ist ein gasförmiger Schadstoff, der im wesentlichen durch die Verbrennung von Kohle, Heizöl und Gas bei der Hausheizung, Stromerzeugung und ähnlichen Produktionsprozessen entsteht. Er verteilt sich im allgemeinen zu einer gleichmäßigen Luftbelastung. Er gefährdet die menschliche Gesundheit. *(optional)*

[Sulfur dioxide is a gaseous pollutant... It is dangerous to human health.]

(7) Der Grenzwert für den Schadstoff Schwefeldioxyd liegt in der Bundesrepublik bei 30 $\mu g/m^3$ Luft für die Langzeitbetrachtung von Durchschnittswerten. Die Kurzzeitbelastung darf nicht höher als 3000 $\mu g/m^3$ liegen (nachzulesen in der TA Luft). *(optional)*

[The threshold value for the pollutant sulfur dioxide is, in Germany, at 30 $\mu g/m^3$ for a long-term observation of average values. The short-term exposition must not be higher than 3000 $\mu g/m^3$ (according to the technical directive "TA Luft").]

Figure 1: A sample target text. The user has chosen from the navigator menus the pollutant SO_2, the location Völklingen, and the period "winter season 1996". In addition, a description of threshold passings was preferred to one of absolute values.

- Report structures must be defined in such a way that their parameters (corresponding to user specifications), their ingredients in terms of assertion patterns, and the required database calls whose results should fill certain places in these patterns are associated to each other in a declarative, flexible, and easily maintainable manner. In particular, common specifications are shared across report types, and data preparation procedures take care of filling specification parameters and data obtained from the database into appropriate places of assertion patterns.
- The instantiation of report structures must be organized in a systematic way, which comprises the selection and the refinement of assertion patterns, depending on relevant database values, lexical material, and context.
- The assertion specifications must be manipulated according to the textual context. Temporal, local, and subject circumstances are not repeated in the presentation.

The second task is the most interesting one in the above list. Assertion specifications originally available in terms of the condensed user parameters ultimately have to be related to lexical specifications in both French and German, the target languages in TEMSIS. In order to achieve these transitions in a systematic way, distinct predicates are defined on three ontological levels, corresponding to:

1. user parameters,
2. conceptual representations,
3. language-neutral representations

There is an increasing degree of explicitness from level 1 to 3. Assume a certain threshold value is included in a report.

1. It is implicitly associated with certain combinations of report specifications, according to a deep analysis of the underlying relations.
2. Its representation on a conceptual level comprises a semantically rich predicate and a value.
3. The concept is expanded into a description at the language-neutral level, distinguishing the threshold from its justification (that is, the law by which it is introduced, and the time period determining its validity).

The techniques of mapping structures across representation levels works on the basis of a small set of compositional schemata, as described in detail in [Horacek, 1996]. The system's functionality within the domain of air quality reporting is well-defined and sufficiently limited for using more condensed mapping schemata than in [Horacek, 1996]. A deep analysis of the underlying relations would be unnecessarily time-consuming in our application.

A fourth ontological level is that of language-specific representation. Here the resulting description may be realized by different word groups in the target language (for instance, 'valeur limite autorisé' and 'gesetzlich zulässiger Grenzwert', respectively, none of the French words corresponding to a single German word and vice versa). This level is implemented through the TG/2 realizer.

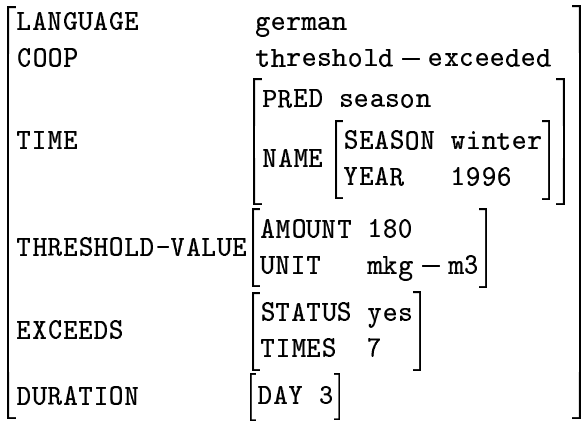

Figure 2: A sample TG/2 input representation for the German version of *During the winter 1996/97, the legally admissible threshold for a three-day assessment of 180 $\mu g/m^3$ was exceeded seven times.*

3 Language-specific realization with TG/2

The system TG/2 [Busemann, 1996; Wein, 1996] is a flexible and reusable, application-oriented text realization system that can be smoothly integrated with deep generation processes. It integrates canned text, templates, and context-free rules into a single production-rule formalism.

TG/2 is based on production system techniques [Davis and King, 1977] that preserve modularity of processing and linguistic knowledge, hence making the system transparent and reusable for various applications.

In the application at hand, the interface between the text planner and TG/2 consists of domain speechact representations. An example is shown in Figure 2. These representations are ignorant with respect to the differences between German and French. Basically, they express a speechact (COOP) combined with a set of domain roles. The roles express e.g. the pollutant, the type and the value of thresholds, actual measurements, information about the time and the place the measurements were taken etc.

The text planner uses this intermediate representation every time it feeds an assertion to TG/2. The planner knows which roles must be expressed for TG/2 to generate output. On the basis of contextual knowledge, it also knows which information should be left out since it is already known to the user. In utterance (4) of Figure 1, the pollutant need not be mentioned, since it was introduced in (3) already. Similarly, utterance (5) need not repeat many parts mentioned already in (4), e.g. details about the threshold, or the location. The text planner decides which types of informations are passed on for realization in TG/2.

Formally, the representations are encoded as feature structures. This allows for a very comfortable adaptation to a TG/2-internal format through the unification with a predefined structure that expresses the necessary mappings as coreferences. The internal representation is looked up by the grammar interpreter in order to determine which production rules can be applied, and what information must be realized into natural language.

Grammar rules are designed as productions (cf. Figure 3). They are encoded in the

```
(defproduction wertueberschreitung "WU06"
  (:PRECOND (:CAT DECL
             :TEST ((pred-eq 'threshold-exceeded)
                    (not (threshold-type-p))))
   :ACTIONS (:TEMPLATE  "Der gesetzlich zulaessige Grenzwert von "
                        (:RULE VAL (get-param 'threshold-value))
                        (:OPTRULE POLL (get-param 'pollutant))
                        "wurde "
                        (:OPTRULE PPtime (get-param 'time))
                        (:RULE DUR (get-param 'duration))
                        (:OPTRULE SITE (get-param 'site))
                        (:RULE EXCEEDS (get-param 'exceeds))
                        ".")))
```

Figure 3: A TGL rule for German encoding a template to be used for input as shown in Figure 2. Information about the threshold value, the duration and about threshold violation are mandatory; all other slots are optional. The function `get-param` extracts the relevant information from the translated input representation.

language TGL [Busemann, 1996]. A rule is applicable, if its preconditions are met. The rule in Figure 3 is applicable to input material as shown in Figure 2, because the `COOP` slot matches, and there is no information about the threshold type (such information would lead to a different sentence pattern). TGL rule development in previous applications showed that it is possible to separate general, linguistic rules from specific ones, thus allowing the general portions to be reused in other applications.[1] The use of different levels of abstraction from underlying message information (canned text, templates, context-free grammars) allows the grammar writer to model general, reusable linguistic knowledge as well as more specific task and domain-oriented wordings. In particular, standardized linguistic realizations of typical situations can be directly encoded into the grammar as canned text.

By associating canned text with domain speech acts, TG/2 behaves in a domain and task specific way. The loss of flexibility in the wording, which the text planner cannot influence, is hardly a problem in technical documents. However, repetition of known information can be avoided. The possibility of omission is reflected in the grammar through the notion of optional rule applications (`OPTRULE`, cf. Figure 3). Optional rules are ignored if the input structure does not contain relevant information. In the domain and for the task at hand, it was possible to design the text templates in such a way that this compositional approach leads to fluent text.

If alternative formulations for some message are encoded in the TGL grammar, they can be ordered according to a set of preference criteria that cause the system to prefer certain formulations over others. Grammar rules leading to preferred formulations are selected first from a conflict set of concurring rules. This way, different formulations can be generated as well as texts of different length. The preference mechanisms will be used to tailor texts for administrative and public uses, respectively.

[1] For instance, a sub-grammar describing dates in the domain of appointment scheduling [Busemann *et al.*, 1997] could be reused here with minor extensions.

4 Conclusion and Future Work

We described ongoing work on the generation of German and French air quality reports on the basis of up-to-date environmental measurements. A prototype implementation has been achieved and is being tested with real data. The texts are generated in either German or French. They are presented as HTML documents (alternatively, LaTeX or plain ASCII text is provided on demand).

We claimed that a simple, compositional approach to text structuring and realization be sufficient for this domain. Obviously there is a large class of relatively simple NLG applications that can be captured by the approach presented in this paper. However, other applications may require some interdependency between the text planner and the realizer, thus calling for a more elaborate and flexible interface including e.g. logical forms for the utterances. While such representations can be handled within TG/2, as was shown in [Busemann, 1996], they require much more effort for grammar development and extension than the flat ones adopted here.

Future work will place particular emphasis on the application-oriented design of the interface between the text structuring component and the realizer. We believe that gathering experience with real applications is a good starting point for determining relations between requirements imposed by the applications and the level of abstraction chosen for the intermediate representations. As a result we will be able to tailor systems better according to the complexity of NLG applications.

References

[Busemann *et al.*, 1997] Stephan Busemann, Thierry Declerck, Abdel Kader Diagne, Luca Dini, Judith Klein, and Sven Schmeier. Natural language dialogue service for appointment scheduling agents. In *Proc. 5th Conference on Applied Natural Language Processing*, pages 25–32, Washington, DC., 1997.

[Busemann, 1996] Stephan Busemann. Best-first surface realization. In Donia Scott, editor, *Eighth International Natural Language Generation Workshop. Proceedings*, Herstmonceux, Univ. of Brighton, England, 1996. Also available as Research Report RR-96-05, Deutsches Forschungszentrum für Künstliche Intelligenz, Saarbrücken, Germany.

[Davis and King, 1977] Randall Davis and Jonathan King. An overview of production systems. In E. W. Elcock and D. Michie, editors, *Machine Intelligence 8*, pages 300–332. Ellis Horwood, Chichester, 1977.

[Horacek, 1996] Helmut Horacek. Lexical choice in expressing metonymic relations in multiple language. *Machine Translation*, (11):109–158, 1996.

[Wein, 1996] Michael Wein. Eine parametrisierbare Generierungskomponente mit generischem Backtracking. Master's thesis, Department for Computer Science, University of the Saarland, 1996.