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

Automatic Production of Multilingual Environmental Information

Bernd Bohnet[1], François Lareau[2], Leo Wanner[2,3]

Abstract

Multilingual environmental information is communicated via different media. These media can be newspapers, TV, internet, WAP, SMS, etc. Each of the media has a presentation mode which fits best. Thus, it turned out that for newspapers, pictograms which indicate good, normal, and bad conditions, a map with pictograms and/or a very short text are to be preferred. In contrast, the information provided in the internet, can be very detailed and personalized and contain the latest data available at the moment the user requests the information. Furthermore, in the internet, the user can interactively select more details or change the presentation mode. For all media, the most challenging information mode is text. Since a template-based method where predefined sentences with empty slots filled at the time of generation cannot ensure coherent and cohesive text for all contextual settings, full-fledged generation techniques are needed. In this paper, we present the generation techniques as used for the production of multilingual air quality information in the framework of the MARQUIS-project.

1. Introduction

To understand something means to get information. Therefore, the data as provided by the measurement stations has to be turned into information. However, it is still standard that the measurements which a lot of stations provide in Europe are just put into tables and curves that only experts can understand. This means that simple citizens cannot grasp the meaning of those numbers, tables, and graphics. This might also be the reason why European legislation already has taken up this topic and made it a law that the citizens must get the environmental data in a form which they can easily understand. As it turns out, text is one of the most easily accessible forms.

For economical reasons, the texts can not be generated manually as there are too many measurement stations producing new data at frequent regular intervals (often hourly). For the communication of environmental information, the technique of automatic information generation is best suited since the texts are not too difficult to produce. At the same time, flexible techniques are needed, since the range of possible situations may be large enough. Moreover, it can be useful to generate the information in different languages. In this paper, we describe the successful realization of such a multilingual information production system for environmental reports. The system was developed during the project MARQUIS (Multimodal air quality information service for general public) founded by the European Union in the eContent program.

The central *information providing features* of the MARQUIS service include: (a) "on the fly" generation of textual information using state of the art Computational Linguistics techniques in Catalan, English, Finnish, French, German, Polish, Portuguese, and Spanish – with region and culture-specific priorities being taken into account; (b) generation of graphics and tables, (c) maintenance of regional profiles with geographical and cultural characteristics as well as of individual customer login profiles concerning the interests, personal characteristics, expertise in the domain, and the information already provided by the service; (d) information delivery on demand and event-driven. More details are found in (Wanner et. al., 2007).

[1] University of Stuttgart, Stuttgart, Germany; e-mail: bernd.bohnet@informatik.uni-stuttgart.de
[2] Universitat Pompeu Fabra; e-mail: francois.lareau@upf.edu
[3] Institucio Catalana de Recerca i Estudis Avançats; e-mail: leo.wanner@icrea.es

In what follows, we start with a description of the requirements as assessed by the linguistic analysis of the domain. Then we describe the process of the linguistic realization.

2. Linguistic Analysis of the Domain

The purpose of linguistic analysis of the domain is to find out the linguistic constructions underlying the information on air quality as to be generated by the MARQUIS-service. The analysis concerns the discourse level as well as the sentence level and was a prerequisite for the work on the development of the document planning module and the sentence generator. It is performed with a corpus and has great impact on the sentence generator since the grammar engineer has to ensure that all relevant information is in the dictionary and the grammar has all rules to build the needed sentences. To illustrate this we give in the next two tables examples of elementary information units which are basic elements of text and circumstantial information units.

2.1 Elementary Information Units

air quality index / air quality	Values of the air quality index (e.g., 'good', 'satisfactory', 'bad', etc.); the number and exact labels of the air quality index values depend on the index used. In MARQUIS, five different country-specific indexes were used.
individual air pollutants	'particulate matter' ('dust particles', 'PM10' / 'PM2.5'), ozone, nitrogen dioxide, sulphur dioxide, etc.
units	$\mu g/m^3$, hour
concentration	e.g. 175 $\mu g/m^3$
thresholds	'information threshold' , 'alarm threshold', etc.
forecasts	'forecast index', 'forecast index', etc.
justification / forecast motivation	weather, traffic, etc.

2.2 Circumstantial Elementary Information Units

Time Circumstantial	day time: 'early morning', 'morning', 'noon', 'afternoon', etc.: time: time point in hours; 'between', 'since', 'from'
Manner Circumstantial	Exceedance, x-times, frequently, nearly; Increasing, decreasing: hardly, slightly, moderately, clearly, etc.
Location Circumstantial	station name, e.g. at the station Ludwigsburg, city name, region, state, country

The above elementary information units are mainly used as concepts. The analysis had also to determine which verbs are used together with the elementary information units. This information was stored in dictionary entries which are used by the grammars. The analysis had to be carried out for all eight languages. The information was then stored in dictionary entries that describe how the words can be combined

in order to form correct sentences. The example below shows the entry for the English word *concentration*:

```
concentration : noun {
  pos = N
  countable = yes
  gp = {
     1 = I
     2 = II
     I = {  dpos = N    rel = compound det = no }
     I = {  dpos = N    rel = noun_completive prep = of  det = no }
     II = { dpos = Num  prep = of }
     II = { dpos = Adj  rel = modificative }
     II = { dpos = Adv  rel = modificative }
  }
  Magn = high
  AntiMagn = low
  Adv1 = in            // "we found ozone in a concentration of 180
µg/m3"
  Func2 = be1          // "the concentration (of ozone) is 180 µg/m3"
  IncepFunc2 = reach   // "the concentration (of ozone) reached 180
µg/m3"
  IncepOper1 = reach   // "ozone will reach a concentration of 180
µg/m3"
}
```

The value of the attribute "I" in the entry means roughly that the word *concentration* needs another word which is a noun. This noun can become a compound (e.g. *ozone concentration*) or a complement (e.g. *concentration of ozone*). The value of "II" means that *concentration* combines with a number, adjective or adverb. Further, the entry specifies verbs which are typical use to build a sentence with this word. If the document plan provides a concentration, a pollutant and value, then this can be combined to sentence like *The ozone concentration is 185 µg/m³*. If the document plan would specify a statement about the concentration, then this information would be used also to put the information in the sentence about the concentration as *The ozone concentration of 185 exceeds the information threshold*.

3. Linguistic realization

The production of multilingual environmental textual information starts with extraction of the relevant information by the assessment module from the data. The assessment module provides the document planner with the relevant information. The document plan represents the environmental information which consists, besides the data itself, of rhetorical relations which give the structure of the document to produce and the presentation mode (text, graphic, table). The micro planner cuts the document plan into pieces using the rhetorical relations. Figure 1 shows a part of such a document plan. It consists of information units like 'AQI' (air quality index), 'very poor', and 'ozone concentration', and rhetorical relations like 'cause' and 'interpretation'. Each of thes pieces becomes a sentence. In the same step, the information which belongs together is grouped, for instance, if the concentration of some pollutants having the same concentration level. Afterwards, the pieces are represented as conceptual graphs (Sowa, 2000) which are still language independent representations. Then a graph transducer (Bohnet, 2006) maps the conceptual graph to semantic representations, after which the semantic representations are mapped to syn-

tactic representations, then further on to topologic/morphologic representations, and finally to sentences. The advantage of this method is its high flexibility. For instance the most appropriate words depending on the data are selected, anaphora and elisions are introduced. Figure 3 shows a overview of the generation process.

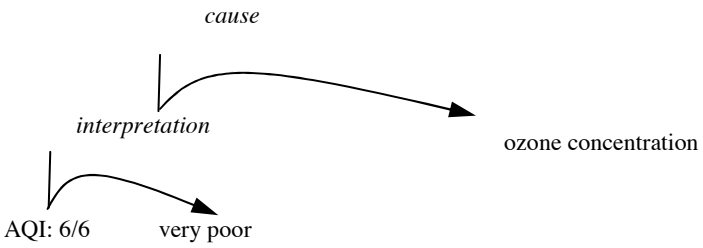


Figure 1: Part of a document Plan

3.1 Conceptualization

The conceptual representations are conceptual graph in terms of Sowa. Figure 2 shows a example. The nodes are labeled with concepts and types. The concepts are concocted by conceptual relations.

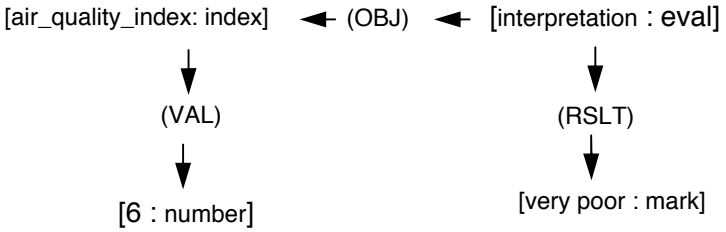
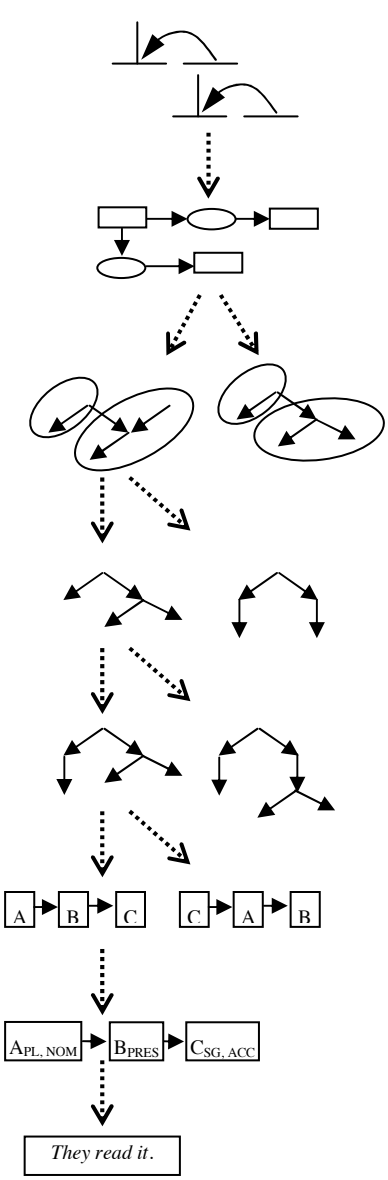


Figure 2: Conceptual Graph



[concentration]
Document plan

Conceptualization
universal concepts
aggregation
referring expressions

Semanticisation
language-dependent meanings
communicative organization

Deep syntacticization
meaningful words
rough sentence structure

Surface syntacticization
grammatical words
fine sentence structure

Linearization
word order
agreement
punctuation

Morphologization
morphological information

Inflection
final form of the words (TLM)
contraction and elision

Figure 3: Overview of the generation process

3.2 Mapping conceptual representations to semantic representations

The conceptual structures are mapped to a semantic structure. A few generic rules define the mapping of concepts to a predicate argument structure. The rules can be kept generic because a dictionary of concepts describes the correspondences between concepts and semantic units. There is a separate dictionary for each language. It stores only the language dependent information, while the generic rules handle the language independent phenomena. Figure 4 shows part of a conceptual dictionary. The first word of an entry is the name of the concept. This name is followed by a colon and the name of concept from which it inherits features.

```
air : T { sem=luft }
air_quality : quality {
      // semantic substructure to which the entry is mapped
     definitness=definite
     // the second argument of the concpet becomes are (German: Luft)
     default = { arg=1 target=luft target_com=theme  }

}

quality : T {
  // semantic substructure to which the entry is mapped.
  sem=qualität // German word for quality
  RSLT = { arg=2     target=referent source_com=theme }
}
```

Figure 4: Entries of a dictionary with concepts and its mapping to predicates and arguments

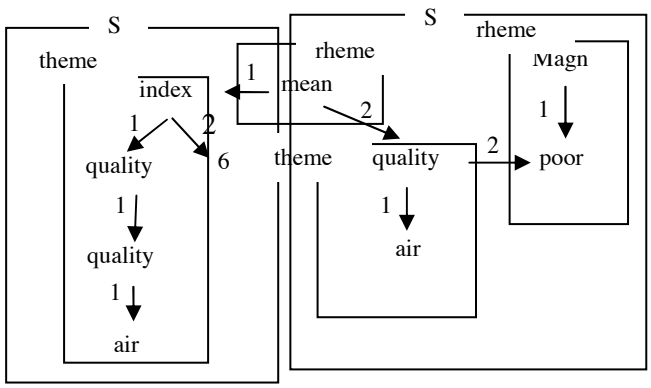


Figure 5: Semantic Representation

The feature *sem* defines the name of the semanteme to which a concept is mapped. For instance, the entry *quality* has the feature *sem=qualität*. Therefore, the German word *qualtität* is chosen as the node label in the semantic structure. Figure 5 shows a semantic representation. The nodes are labeled with the predicates and arguments. The number of the edge give the number of the argument. The rectangles indicates sentences and the communicative structure of the sentence.

3.3 Deep Syntacticization

The deep syntactic representation is a tree which contains words, that have meaning, i.e., the function words like most of the prepositions, light verbs and determiners are not contained. The mapping is driven by elementary rules that typically handle only one or two edges. Therefore, most of the rules are general and easily combinable. The rules use additional information from a typed feature structure which is a tree. The structure contains information of the single words and their behavior, i.e., the lexical category of a predicate / argument, the syntactic relation between semantic units, etc.

3.4 Surface Syntacticization

The surface syntactic representation is a tree which contains all the words of a sentence. The edges are labeled with surface syntactic relation like subject, direct object, determiner, etc. In this step, the grammar has to add function words (preposition, light verbs, determiner, etc.) and to labeled the edges with the surface syntactic relations. Again a lot of information about the behavior of the words come from a typed feature structure which is considered by the rules.

3.5 Linearization

The topologic representation is a hierarchical graph. It consists of word order domains and precedence relations. Each word order domain builds on a surface sequence of words that means that no word from outside of a domain can go between words of a domain. The domains are recursively (hierarchically) organized. The precedence relation order words and domains. The actual word order is computed by a topologic sort. In this step, morphologic rules derive the morphologic features.

3.6 Morphologization and Inflection

This representation is just a chain of strings. The strings are computed out of the word forms and the morphologic features. The features are just attached to the basic word form. In the Marquis-project, a Two-Level-Morphology model computes the inflected word form.

3.7 Example

The following text shows an example.

The air quality index is 3, which means that the air quality is satisfactory. This is due to the ozone concentration. The nitrogen dioxide concentration and the PM10 concentration do not have influence on the index. The current air quality index (3) is the highest. The lowest air quality index was 2 (at midnight). Between midnight and 9 AM, the air quality index remained stable at 2 and between 10 AM and 3 PM, it remained stable at 3.

The ozone concentration (70 μg/m^3) is relatively low. As a result, no harmful effects to human health are expected. Between 4 AM and 10 AM, the ozone concentration increased considerably from 22 to 76. The current ozone concentration (70 μg/m^3) is close to the highest of 76 μg/m^3 (at 10 AM). The lowest was 22 μg/m^3 (at 4 AM).

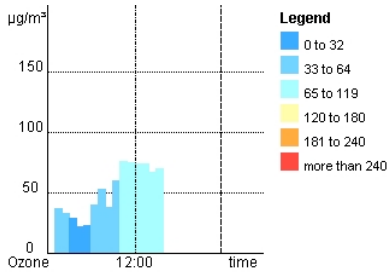


Figure 6: A sample English text

Tomorrow, the ozone concentration (45 μg/m^3) will be low. This is due to the bad weather.

El índice de calidad del aire es 3, lo que significa que la calidad del aire es aceptable. Esto es debido a la concentración de ozono. La concentración de dióxido de nitrógeno y la concentración de PM10 no contribuyen al índice. El índice de calidad del aire actual (3) es el máximo. El índice de calidad del aire mínimo era 2 (a las 0:00). Entre las 0:00 y las 9:00, el índice de calidad del aire se mantuvo estable en 2 y entre las 10:00 y las 18:00, permaneció estable en 3.

Figure 7: A Spanish text for the same station, but some hours later

4. Conclusions

We have build a system for the automatic production of multilingual environmental information. The system creates reports about the current situation of several air pollutants like ozone, PM10, sulfur dioxide, carbon monoxide in eight languages. It is based on a full fledged text generator. Within the time of the project, we have build reusable linguistic resource for environmental reports.

Bibliography

Bohnet, B. (2006): Textgenerierung durch Transduktion linguistischer Strukturen. DISKI 298, Köln.

Mel'cuk, I.A. (1988): Dependency Syntax. State University of New York Press. Albany. NY.

Sowa, J. (2000): Knowledge Representation, Pacific Grove

Wanner L., (et al) (2007): From Measurement Data to Environmental nformation - 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.