Automatic Generation of Instructions in a Multilingual Environment ⋆

Ivana Kruijff-Korbayová

Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
korbay@ufal.mff.cuni.cz

Abstract. The paper describes work that is done in the context of the an international project called AGILE (Automatic Generation of Instructions in Languages of Eastern Europe). The overall aim of the project is to develop a Multilingual system for generating continuous instructional texts in Bulgarian, Czech and Russian. The project is concerned mainly with (i) the development or adaptation of linguistic resources for the chosen languages, and (ii) the investigation and specification of text structuring strategies employed in those languages for the given type of texts.

1 The AGILE Project

The primary goal of the AGILE project is to develop a suite of software tools to assist technical writers in the production of software manuals in selected languages of Eastern Europe, namely Bulgarian, Czech and Russian. The approach chosen in this project is *multilingual generation* from a common semantic representation of the procedural aspects of the tasks involved in using given software tools. The distinctive feature of language generation, as compared to machine translation, is that the meaning of the text is encoded in a formal language rather than a(nother) natural language. The formal language needs to be precise, so that problems of ambiguity are avoided.

In AGILE, one set of the tools being developed enables the software developer to specify the end users' domain tasks and the operations at the level of the user interface. It relates interface actions (e.g., mouse clicking, menu opening), interface objects (e.g., icons, windows) and application actions (e.g., modify parameter) to domain tasks (e.g., creating a wire-frame image). The result is a specification of the content of the instructional material to be generated.

The second set of tools assists the technical writer by transforming the constructed domain task model into draft documents in Bulgarian, Czech and Russian. Multilingual documentation is thus generated directly from the user interface and domain task model, in contrast to the current practice, where the initial documentation is written in one language only and is subsequently translated.

⋆ The research was carried out in the project AGILE started in January 1998 and funded within the EU programme INCO-COPERNICUS, grant No. PL96114. This paper describes the results of joint work of the Charles University team. I would like to thank especially Jiří Hana and Hana Skoumalová for their contribution.

An important feature of the approach followed in the AGILE project is that we address not only the linguistic phenomena within a sentence, but concentrate on higher-level textual and pragmatic issues, in particular the interdependence between discourse and syntactic choices.

The architecture of the system for text generation developed in AGILE is based on the experience gained in the development of the DRAFTER system for English (Pemberton et al. 1996). The system is intended for use in a scenario which involves a user, i.e. a technical writer or a a software designer producing a CAD-CAM software user guide or manual, on the basis of his/her knowledge of the software product, esp. its functionality and user interface. The user defines the content of the texts through a user interface tailored to the CAD-CAM domain, and then the system generates the corresponding texts in Bulgarian, Czech and Russian automatically. The process of generating a text in AGILE thus consists of the following steps:

1. the user specifies the contents of the text to be generated as a domain task model, common to all three languages
2. the text structuring component chooses a text template for the text to be generated and creates a text plan which specifies in what order and by what means the content is to be conveyed in a sequence of sentences; the resulting text plan can be the same for all three languages, or different plans can reflect differences among languages at the discourse level; each sentence is specified by a set of constraints in an expression in the form of SPL (sentence planning language) (Kasper 1989)
3. the tactical generators process each SPL to produce the sequence of sentences in each of the three languages conveying the given content according to the constraints imposed by the text planner; the tactical generators exploit the language-specific components, i.e. grammars, lexica and morphology components
4. if the chosen text template includes layout and format requirements, these are applied to the generated sentences in order to obtain an appropriate output text

Let us look at each of the steps in more detail in the following sections.

2 Domain Model

The contents of a text to be generated in AGILE are specified in a formal language as the so-called domain task model. Using knowledge representation terminology, the domain task model is an A(ssertion)-box, which instantiates concepts defined in the T(erminology)-box. The T-box constitutes the domain model, i.e., a set of concepts representing knowledge in a specific domain –in the case of AGILE the domain of CAD/CAM applications. It determines the range of meanings that can be exploited in a particular domain task model, and thus expressed in a text. The T-box has a general and a specific part. The general part is called the "Upper Model", and includes abstract notions like "quality" and "process" which are linked to syntactic features of natural language (roughly, adjective and verb). These concepts are abstract and universal, and therefore they would be applicable to any language and any subject-matter. The specific part of the T-box comprises a set of concepts for modelling a set of CAD/CAM applications (e.g., "multiline", "justification"), which is the subject-matter chosen for AGILE. By linking these specific concepts to the abstract Upper Model concepts, we encode constraints on how they may be expressed linguistically.

Most of the AGILE T-box has been drawn from two existing resources: the Penman Upper Model (Bateman et al. 1990), and the concepts for modelling instructions used in the DRAFTER project (Pemberton et al. 1996). To these some concepts specifically relevant to the CAD/CAM domain have been added. For reasons of simplicity and efficiency, the domain model has been implemented in the Common Lisp Object System CLOS.[1]

3 Text Structuring

Taking into account the current potential of the natural language generation technology in general, one cannot at present aim to generate texts with wide-ranging conceptual content or advanced stylistic features. Rather, the texts considered for generation by the system developed in AGILE first of all correspond to routine passages, which occur frequently in the user guides and manuals, have a quite simple semantic content and have a regular style, terminology and structure (Power and Scott 1997). As expected on the basis of the experience in DRAFTER, the preliminary text analysis of the user guide of the AutoCAD system revealed, that these criteria are satisfied by descriptions of step-by-step procedures for performing a task. A corpus of such texts has been assembled and studied, as reported in (Hartley et al. 1998a). On the basis of this corpus, a set of simple text structures has been identified. These were chosen to be generated for the purpose of the initial demonstrator.

The task of the text structuring module is to take the user's definitions of the text content as an input, and to yield a set of SPLs which determine how this content is to be conveyed in each of the languages in a text in a given style. This task involves macro-planning of the text, the splitting up of the content into chunks that can be expressed by sentences and also determination of the particular ways of expressing the content, depending on the surrounding context. These are all complex issues, which we have to handle in an incremental fashion in the successive stages of the project.

In the first phase of the AGILE project, simple text structures have been generated for which a quite simple strategy suffices. However, in order to be extendible in the next phases of the project, the strategy needs to be sufficiently general and flexible. We decided to employ text templates corresponding to particular text styles as the guidance in text planning. Each text template has pre-defined components which are filled by content from particular parts of the

[1] The current state of the AGILE domain model is described in (Power 1998).

user's definition of the text content, and for which the typical ways of expressing the content in a given language are specified for the template. Furthermore, one can associate layout rules with each component of the text template, so that the desired formatting and type-setting of the generated text can be obtained when the text is printed or displayed as the output of the system.[2]

4 Tactical Generation

The platform for development and implementation of the grammatical resources of Bulgarian, Czech and Russian chosen in AGILE is the Komet-Penman Multi-Lingual (KPML) system (Bateman 1997). The linguistic-theoretical basis of the KPML system is Systemic Functional Linguistics (SFL, Halliday 1985). SFL is characterized by the notions of function and system. SFL is functional in that it acknowledges three broad functions languages have: the ideational, the interpersonal and the textual. SFL is systemic in that the main focus in description is on the grammatical paradigm (or: system). The kernel of SFL is Systemic Functional Grammar (SFG). The grammar of a language is represented as a system network, which can be read as a declarative statement of grammatical features and the co-occurrence constraints holding between them. SFG thus employs a classification-based approach to grammar, rather than a rule-based one.

Adopting a systemic functional model as a linguistic basis for generation generally supports the modelling of tasks involved in generation because many of the characteristics of the systemic functional model are suggestive of the kinds of resources needed in natural language generation. This is evidenced by the fact that many generation systems are based on or have been inspired by SFL, and many projects have used the SFG-based tactical generators Penman and KPML.[3]

The tactical generation process in the KPML generator starts with an input SPL expression (Kasper 1989). Two examples sentences differing in mood and the SPLs that ensure their generation are shown in (1) and (2).

(1) Zadejte koncový bod čáry. (Enter the end-point of the line.)

```
(EXAMPLE
   :NAME   DO-TEXT1-Cz-3-imp   :SET-NAME   DB-TEXT1
   :TARGETFORM   " Zadejte koncovybod cary."
   :LOGICALFORM
   (E / DIRECTED-ACTION
      :LEX ZADAT :SPEECHACT IMPERATIVE :ACTEE (D / OBJECT
         :LEX BOD :ORDINATION-Q ORDINATIVE :ORDINATION-ID
         (O1 / THING
            :ORDINAL-TYPE-Q LAST
            :ORDINATIVE-DETERMINATION-Q NOTSPECIFIED)
         :ORDINATIVE-SELECTION-Q NONSELECTION
```

[2] The current state of the AGILE text structurer is described in (Bateman et al. 1998).
[3] For a detailed list of references to such systems see (Bateman et al. 1998b).

```
          :IDENTIFIABILITY-Q IDENTIFIABLE
          :PART-OF (P / OBJECT
             :LEX CARA :IDENTIFIABILITY-Q IDENTIFIABLE)))))
```

(2) Zadáme koncový bod čáry. (We enter the end-point of the line.)

```
(EXAMPLE
   :NAME   DO-TEXT1-Cz-3-decl   :SET-NAME   DB-TEXT1
   :TARGETFORM   " Zadame koncovy bod cary."
   :LOGICALFORM
   (E / DIRECTED-ACTION
      :LEX ZADAT  :PROPOSAL-Q proposal
      :PERCEPTIVE-REACTIVE-PHENOMENON-Q PHENOMENON
      :ACTOR (SP / PERSON  :NUMBER PLURAL)
      :ACTEE (D / OBJECT
         :LEX BOD  :ORDINATION-Q ORDINATIVE  :ORDINATION-ID
         (O1 / THING
            :ORDINAL-TYPE-Q LAST
            :ORDINATIVE-DETERMINATION-Q NOTSPECIFIED)
         :ORDINATIVE-SELECTION-Q NONSELECTION
         :IDENTIFIABILITY-Q IDENTIFIABLE
         :PART-OF (P / OBJECT
            :LEX CARA :IDENTIFIABILITY-Q IDENTIFIABLE)))))
```

The SPLs are instances of the Upper Model (Bateman et al. 1990; Henschel and Bateman 1994), which is the resort of ideational (i.e., propositional) meaning in KPML. Besides the ideational information expressed in an SPL, also interpersonal information (e.g., `:SPEECHACT command`) and textual information (e.g., `:identifiability-q identifiable`, `:theme o1`) is contained. The non-ideational kinds of information typically come from text planning (see Section 3).

Given the SPL as input, the traversal of the system network is started. At each choice point (each system), a chooser is invoked which poses its inquiries to the semantics (i.e., it checks the input SPL for information enabling a decision to be made). A successful traversal of the system network yields a sentence in the given language.

Multilinguality is realized by resource sharing in KPML. Given that the three languages we are dealing with in AGILE have many commonalities, there is no need to develop three separate grammars, but rather, we share as much of the system networks as possible and only develop separate components reflecting differences among the languages.

5 Concluding Remarks

Due to the limited space of this article, the system for multilingual generation developed in the AGILE project could be described only briefly. The main components and the essential methodology have been sketched. In the current phase of the project, we are able to generate texts involving a number of concepts from the CAD/CAM domain. These texts have a simple structure and contain essential grammatical constructions used in instructions. In further phases of the project we will include textual and syntactic phenomena of higher complexity.

One of the interesting challenges in the course of developing the grammars for Czech, Russian and partially also Bulgarian is their high degree of free word order and the fact that word order is determined mainly contextually. The approach chosen in AGILE enables us to tackle this issue, since the interaction between discourse and syntax is taken into account.

A number of interesting issues arise in the area of text planning. One issue is a contextually appropriate choice of forms of referring expressions, e.g., the decision how much descriptive content needs to be included. Another issue is a contextually appropriate choice between a sequence of sentences and a complex sentence with respect to the complexity of the content expressed as well as the complexity of the generated output.

References

J.A. Bateman, R. Kasper, J. D. Moore, and R. Whitney. A general organization of knowledge for natural language processing: the Penman Upper Model. Technical report, USC/ISI, Marina del Rey, California, 1990.

J. Bateman, A. Hartley, I. Kruijff-Korbayová, D. Dochev, N. Gromova, K. Staykova, A. Bémová, A. Rosen, J. Trojánek, and E. Sokolova. Tagging and analysis of instructional texts in the software domain. AGILE project deliverable WP3: CORP, CORP-Bu, CORP-Cz, CORP-Ru, University of Brighton, UK, 1998.

J. Bateman, E. Teich, D. Dochev, N. Gromova, J. Hana, I. Kruijff-Korbayová, S. Sharoff, and E. Sokolova. Grammatical resource implementation for Bulgarian, Czech and Russian. AGILE project deliverable WP7: IMPL1-Bu, IMPL1-Cz, IMPL1-Ru, University of Brighton, UK, 1998.

J. A. Bateman. Enabling technology for multilingual natural language generation: the KPML environment. *Natural Language Engineering*, 1(1), 1997.

M.A.K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London, 1985.

A. Hartley, D. Dochev, N. Gromova, J. Hana, S. Sharoff, and E. Sokolova. Generation of simple text structures for instructions in Bulgarian, Czech and Russian. Agile project deliverable WP5: TEXS1-Bu, TEXS1-Cz, TEXS1-Ru, TEXM1, University of Brighton, UK, 1998.

R. Henschel and J. A. Bateman. The merged upper model: a linguistic ontology for German and English. In *Proceedings of COLING'94, Kyoto, Japan, August 1994*, 1994.

R. T. Kasper. A flexible interface for linking applications to PENMAN's sentence generator. In *Proceedings of the DARPA Workshop on Speech and Natural Language*, 1989.

L. Pemberton, L. Gorman, A. Hartley, and R. Power. Computer support for producing software documentation: Some possible futures. In T. van der Geest and M. Sharples, editors, *The New Writing Environment: Writers at Work in a World of Technology*. Springer Verlag, 1996.

R. Power. Preliminary model of the CAD/CAM domain. AGILE project deliverable WP2: MODL1, University of Brighton, UK, 1998.

R. Power and D. Scott. NLG tools to support technical authors and translators. A manuscript submitted to the ANLP'97 conference, 1997.