Proc. of the Ninth Int. Conference on Tools with Artificial Intelligence (TAI'97),
November, 3-8, 1997, Newport Beach, California, pp.182-189

# A User-Assisted Business Letter Generator Dealing with Text's Stylistic Variations

E. Stamatatos, S. Michos, N. Fakotakis and G. Kokkinakis
*Dept. of Electrical and Computer Engineering*
*University of Patras*
*stamatatos@wcl.ee.upatras.gr*

## Abstract

*This paper describes a user-assisted business letter generator that meets the ever-increasing demand for more flexible and modular letter generators which draw on explicit thematic models and are easily adaptable to specific user needs. Based on a detailed analysis of requirements and taking full advantage of the end users feedback, the presented generator not only creates a business letter according to the user choices, but also refines it taking into consideration stylistic aspects like written style and tone.*

## 1. Motivation

Correspondence is essential in establishing and confirming transactions in commerce. Although there are several competing forms of communications in the modern world (e.g., telephone, fax, telex, e-mail), an old-fashioned letter is still the most formal way to get the message across, since correspondence that's prompt, accurate, and professional looking, creates a sense of responsiveness, dependability, and trust [10].

The creation of business letters is an important activity in office environment applications. Taking into consideration the large amount of such common and stereotyped letters produced daily in an office, it is easy to understand that considerable losses in productivity as well as fruitless wastes in time and money are observed for such a trivial, but necessary, activity. Therefore, it is obvious that the automation of this procedure would help companies achieve better correspondence quality and avoid a time-consuming and expensive task.

The vast majority of business letters is a short (up to one page) standardized text. In many cases, similar letters must be produced several times with slight alterations. For any business a finite number of pre-written letters and paragraphs can cover virtually all situations. Additionally, a business letter often has to be composed in two or more languages [2]. Therefore, providing this possibility automatically could considerably facilitate the task of composing a letter, particularly for a non-native speaker of a language.

Moreover, commercial correspondence style obeys to some specific rules. There are specific expressions that set the tone of the letter and modify the addressee's reactions. Business letters are often written in an old-fashioned, pompous style which usually complicates the message [2]. Hence, setting automatically the interpersonal and situation aspects that affect the form and content of a business letter would help in establishing some simple and clear stylistic variations in a text that will convey the useful information without misunderstandings.

This paper presents an approach to the automation of business letter composition emphasizing on the above points. The next section contains a description of relevant work in natural language generation (NLG). Section 3 comprises the results of the analysis of a medium-sized business letter corpus and section 4 describes our approach in detail. Section 5 includes two examples of the system output together with some comments. Finally, in section 6 a brief discussion on system's performance is given and some conclusions are drawn.

## 2. Relevant work

During the last twenty years research in NLG has managed to offer robust general-purpose single-sentence generators and limited-purpose multisentence generators [8]. Some fruitful areas for application have been investigated. Particular promising areas seem to be situations where data change quickly, or where large amounts of data have to be extracted from a knowledge base and presented in a standardized way. So, the most popular and successful NLG applications to date are the generation of stock market reports [9], weather forecasts [6], technical reports [13] and business letters [14], [5].

Text generation as it is used in the vast majority of up-to-date software is limited to **canned text**. This technique, as opposed to linguistically founded text generation, has the following shortcomings:

- the range of texts and the stylistic variations these systems can produce are very limited;
- the texts are not context sensitive. They are not tailored to the user. Therefore, the information conveyed or the form used may be inadequate for a specific user;
- extending and updating these systems is hard, if not impossible.

**Template systems** are used as soon as a message must be produced several times with slight alterations. They are the most robust approach for generating multisentence texts [9]. Form letters are a typical template application, in which a few open fields are filled in specified constrained ways. More sophisticated template systems fill template slots with linguistically-generated fragments [13]. In such systems, low-level phenomena (i.e., agreement, morphology, punctuation reduction, etc.) as well as more sophisticated ones (i.e., ellipsis, relative-clause introduction, etc.) cannot be easily handled. Moreover, if the number of templates becomes large, then it is hard to maintain and update these systems. Nevertheless, it seems that they are the most suitable approach for an application that requires the composition of standardized texts.

Linguistically based systems such as **phrase-based systems** [7] and **feature-based systems** [3] are not yet able to operate beyond the experimental level, especially for generating multisentence texts. These systems produce higher-quality output but building them is an expensive task that requires a large amount of computational effort. Additionally, they need some kind of representation of the information included in the text [6]. Such representations do not exist in the great majority of today's application systems.

On the other hand, style is the main factor besides the propositional content that modifies the listener's reactions. There are some approaches that try to take advantage of **stylistic aspects** in order to improve the quality of a text generator output [7]. Though stylistic aspects are too vague in their nature, it is obvious that stylistic variations depending on the topic of the text, the interlocutor, and the communication instance, are crucial for giving the reader the sense that (s)he doesn't read a machine-generated text. The main problem is that it is not yet possible to obtain a formal description of style that could be used in both style identification in text understanding and style generation in NLG. Furthermore, it is not yet clear how we can apply the results of our previous work in style identification to NLG [11]. Previous approaches to automatic generation of commercial correspondence did not pay special attention to stylistic variations of the produced texts, despite the fact that style is considered by them to be one of the most important factors for producing high quality letters [14].

Finally, the development of text plan libraries, that is representational paradigms for characterizing stereotypical texts, seems to be the crucial point in NLG research [8]. Such libraries would lead to reliable text planners capable of coping with real-world domains and planing texts of several paragraphs. So far no such text planner exists.

For all the above reasons, we decided to utilize template-based generation in the composition of business letters and take advantage of stylistic aspects, like written style and tone, in order to improve the quality of our text generator output. In the next sections, we will see how close to this initial goal are the output results.

## 3. Analysis of business letters

### 3.1. Sections of a business letter body

It is common ground that the quality of the output of a text generator depends on how well the text parameters (i.e., language, theme, written style, tone, etc.) have been analyzed and taken into consideration during its implementation phase. Towards this end, we analyzed manually a corpus of business letters that were collected from companies, organizations, institutions, and books written by experts on this field. This medium-sized corpus (about 1500 business letters) comprises common letters in Greek referring mainly to commercial transactions and announcements.

A business letter is composed of several fields: *the addresser's name, the addressee's name, the current date and place, a salutation, the body of the letter, a complimentary close, a list of enclosures* (optional), etc. [4]. In this work we deal with the **generation of the body of a letter**, that is the field conveying the important information.

In general, a well-formed business letter consists of three main sections [2], [10], [4]:

1. Cause: a statement of why the addresser was prompted to edit the letter. It is usually located at the beginning of the letter (i.e., the first paragraph). For instance, the addresser may want to reply to a previous letter sent by the correspondent.
2. Purpose: it concerns the basic points that need to be stated. For instance, the addresser may want to place an order for a product.
3. Conclusions: it comprises wishes and proposals which the addresser usually ends the letter with (i.e., the last paragraph).

Each one of the above sections consists of a set of *components*, that is the primary conceptual elements that compose the letter [4]. For example, the *conclusions* section of a letter may include the components: *further_inquiry_encouragement, thanks_for_writing,* etc. It has to be underlined here that a concrete component

may be common in two or more letters belonging to different categories.

### 3.2. Extraction of letter plans

In order to extract general models for business letters, that is **letter plans**, we used the aforementioned business letter corpus. The first step was **the classification of this corpus into thematic categories**. Then, we focused on the most important categories. In particular, we chose to deal with the following general thematic categories:

- purchase and sale of products: it comprises catalogues, price-lists, samples, inquiries and replies, orders, complaints and adjustments on an order, etc.
- quotations: it comprises quotation of products, new products announcements, etc.
- payment settlements: it comprises invoices and statements, advice and acknowledgment of payment, etc.

The above categories were selected as they cover an important portion of the business letters spectrum [2]. In particular, over 80% of the aforementioned business letter corpus contained letters of these categories. Moreover, the letters that belong to these categories are stereotypical and it is possible to extract simple models that encode their structure.

These categories can be subcategorized further into **primitive thematic categories** that correspond to a specific theme. In other words, these sub-categories cover all the possible instances of the main category (as it was previously shown). For instance, concerning the *purchase_and_sale_of_products* category: consider that company-1 inquires catalogues, price-lists, etc., from company-2 (i.e., *inquiring_catalogues* letters) which, in turn, could reply to company-1 by sending catalogues, price-lists, etc., (i.e., *replying_inquiry* letters). Then, company-1 may want to place an order (i.e., *placing_an_order* letters) and company-2 either acknowledges this order (i.e., *acknowledging_an_order* letters) or refuses it (i.e., *refusing_an_order* letters), and so on [4].

The next step of the analysis of the corpus was **the extraction of the letter plans** that represent the structure of every primitive thematic category. Such a model describes what components may be included in each one of the aforementioned letter sections, what the correct order of these components is, and whether they are mandatory or not. In other words, a model represents a characteristic paradigm of a primitive thematic category that is composed of all the possible components in the right order. The set of these models forms our letter plan library and is language-independent since it is based on conceptual entities [12]. The initial structure we used in Prolog for a letter plan is:

*plan(Thematic_Category,*
*[new_section,component(CodeNo,TYPE),...])*

where the constant *new_section* states the starting of a new section (i.e., normally, each letter plan has three sections), and *TYPE* defines whether the component can be placed in a sentence containing other components or has to be placed in its own sentence (see the description of the sentence planning module in section 4.2).

### 3.3. Stylistic analysis

We processed the aforementioned corpus linguistically (i.e., by hand) in order to **identify all the possible linguistic expressions**, that is, canned phrases or sentences, that fit each one of the components of a section of a letter. It was found that each component may be generated by certain phrases. Furthermore, a sentence may consist of several phrases that correspond to several components. It was observed, however, that in the most formal and well-written letters each sentence included up to two components. In particular, about 60% of the sentences of these letters included two components and over 30% of the sentences included only one component.

Finally, the linguistic expressions extracted with this process were analyzed stylistically (i.e., by hand), in order to **identify their stylistic impact** on the letter. Since business letters is a restricted domain of texts, the stylistic variations were not clear. Nevertheless, slight alterations in terms of written style and tone have been detected. Hence, each phrase that generates a component has been analyzed in two dimensions:

- written style: *elegant*, *typical* and *official*
- tone: *friendly*, *usual*, *severe* and *offensive*

The written style of a text is characterized in a high degree by a set of words/phrases, that is mainly adjectives, adverbs, adverbial phrases, and prepositional phrases [11]. We call these words/phrases **style modulators**, and they can usually be removed from the text without the lack of any important propositional information. So, the more style modulators a phrase includes, the more elegant the phrase is. Furthermore, there are certain stereotypical phrases that set the letter tone as friendly, usual, severe and offensive. This can be seen easily in *complaints* letters where the addresser's tone may vary from very friendly to rude. For example, consider the stylistic distinctions between the following phrases taken from *payment_settlement* letters:

*i)* *we will sue you if your debt is not cleared within the next ten days* (offensive tone)

*ii)* *we are reluctant to place the matter in the hands of solicitors and are offering you a further ten days to settle the account* (friendly tone)

Certainly, there are thematic categories that such tone variations are of no practical use. For instance, in *placing_an_order* letters the addresser can't be rude [2].

## 4. Our approach

### 4.1. Basic characteristics of the system

During the design phase of our generator the user requirements as well as the types of reader characteristics and goals were taken into account. The presented generator is a **semi-automatic multisentence** generator that is based on templates. Template slots are filled by linguistically-processed fragments. In other words, the user determines the content of the letter by selecting the components that will be included in it and by filling in the appropriate templates, where it is necessary. This procedure is simplified via a simple and **user-friendly** interface designed and implemented in Visual Prolog that guides the user and helps him obtain the letter (s)he wishes for.

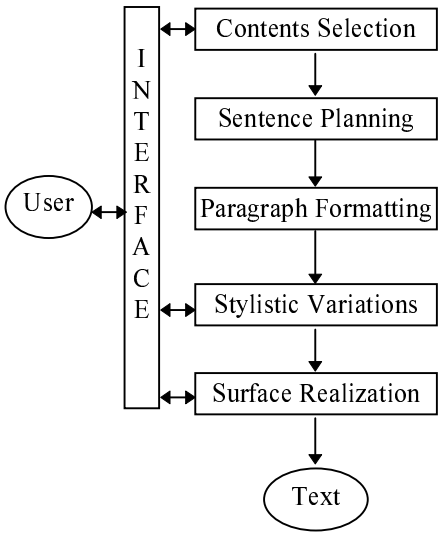


**Figure 1. Overview of the generator**

Although our system currently generates Greek business letters, the generation procedure has been designed with the view of achieving as much **language-independent** results as possible. Hence, our generator could be easily extended to another language (e.g., English) with least additional effort.

Moreover, the presented approach is based on a **pseudo-non-deterministic** generation procedure, that is, the same input can lead to different outputs. Of course, these outputs will have the same characteristics regarding the content of the letter, but they may differ either in the linguistic expressions used in the letter, or in the stylistic variations. Regarding these **stylistic variations**, the user is able to set the desired kind of written style and tone for the letter.

An overview of our generator is depicted in figure 1. The following paragraph describes analytically the generator modules.

### 4.2. Description of the system modules

The function of the system modules is as follows:

- **Contents selection**

As it has been mentioned earlier, each primitive thematic category possess a conceptual model that is composed of a **set of ordered components**. Additionally, a component may be **mandatory** or **optional**. So, after the selection of a thematic category by the user, the set of components that constitute the corresponding model is available and the user can select the optional component(s) that will be included in the letter. Hence, the user determines the contents of the produced letter according to his requests. Furthermore, the user also determines the size of the letter, since the less selected components the shorter the letter is. The internal structure we used for representing a component in Prolog is:

*component(CodeNo,[phrase(PHRASE,TONE),...])*

where *PHRASE* is a canned phrase string containing some possible open template slots together with some possible style modulators and *TONE* is the tone category of the phrase.

- **Sentence planning**

The next step in the generation procedure is the determination of the contents of each sentence, that is, the components that form each sentence. As it has been already noted, in a well-written letter each sentence includes up to two components. Additionally, there are certain components that have to be placed in their own sentence (e.g., for the sake of emphasis). Taking all these into account, the selected components are formed so as each sentence includes two components, if possible, or only one, otherwise. This procedure is performed automatically (i.e., it requires no interaction with the user).

- **Paragraph formatting**

Normally, the body of the letter is composed of three paragraphs, that is, one for each section (i.e., cause, purpose, conclusions). Nevertheless, according to the selected components and the sentence planning, two or even all three paragraphs may be unified into one. Additionally, if a section contains only one selected component, it can be included in the paragraph of the

**Example Output 1**

Stylistic Settings: Written Style = *elegant*, Tone = *friendly*

Κύριε,

Σας γνωρίζουμε ότι δεν έχουμε παραλάβει ακόμη την υπ' αριθμόν 107/1997 παραγγελία μας αν και ως κανονική ημερομηνία παραλαβής των εμπορευμάτων είχε οριστεί η 21/3/1997.

Σας παρακαλούμε να ενεργήσετε για την άμεση παράδοση όλων των ειδών που αναφέρονται στην παραγγελία μας. Παρακαλούμε επίσης να μας γνωρίσετε το συντομότερο αν είστε σε θέση να εκτελέσετε την παραγγελία μας, διαφορετικά θα βρεθούμε στην δυσάρεστη θέση να ακυρώσουμε την συμφωνία μας.

Περιμένουμε νέα σας.

Με ιδιαίτερη τιμή

**Translation:**

Dear Sir,

We would like to inform you that we have not yet to date received our order No. 107/1997 though the regular date of delivery of our merchandise was the 21/3/1997.

Would you please take the necessary steps so that all the items listed in our order be delivered immediately. Please let us know as soon as possible if you are in a position to deliver our order, otherwise to our regret we will have to cancel our agreement.

We look forward to hearing from you.

Yours sincerely

**Figure 2. An example of the system output: *friendly tone*.**

previous section. Hence, the produced letter may be consisted by less than three paragraphs. This procedure is also performed automatically, that is, it requests no interaction with the user as it is shown in figure 1.

- **Stylistic variations**

The user is able to select the tone (s)he wishes for the letter among three alternatives: friendly, usual, severe or offensive. For each component there is an appropriate set of phrases that realize it. Each phrase of this set belongs to one of the above tone categories. Hence, by setting the tone of the letter the user selects a subset of phrases that can realize each component. Furthermore, as it has already been mentioned style modulators contained in a phrase determine in a high degree how elegant the letter is. The user can set the degree of elegance for the phrases belonging to the selected tone category by selecting one of the following prompted alternatives:

- elegant: the maximum possible set of style modulators are inserted to the phrases.
- typical: a normal number style modulators is inserted to the phrases (up to one adjective or adverb for every phrase, if possible).
- official: no style modulators are inserted to the phrases.

- **Surface realization**

Each component is generated by a phrase that may be either a simple canned phrase or a canned phrase with a template slot. In the latter case, the user has to fill in properly the template. The surface realization module processes the user's input in order to form as linguistically-correct phrases as possible. For instance, if

**Example Output 2**

Stylistic Settings: Written Style = *elegant*, Tone = *offensive*

Κύριε,

Σας γνωρίζουμε ότι βρισκόμαστε σε εξαιρετικά δύσκολη θέση εξαιτίας σας δεδομένου ότι δεν έχουμε παραλάβει ακόμη την υπ' αριθμόν 107/1997 παραγγελία μας παρά τις διαβεβαιώσεις σας ότι θα είχαμε στα χέρια μας τα εμπορεύματα όχι αργότερα από την 21/3/1997.

Απαιτούμε την άμεση παράδοση όλων των ειδών που αναφέρονται στην παραγγελία μας. Σε περίπτωση που δεν είστε σε θέση να ανταποκριθείτε στις ανειλημένες υποχρεώσεις σας ειδοποιήστε μας σχετικώς, διαφορετικά θα μας αναγκάσετε να ακυρώσουμε την συμφωνία μας.

Περιμένουμε απάντησή σας το συντομότερο δυνατόν.

Με τιμή

**Translation:**

Dear Sir,

We would like to inform you that we are in an extremely difficult position because you have not delivered our order No. 107/1997 despite your assurances that we would receive our merchandise no later than 21/3/1997.

We demand immediate delivery of all the items listed in our order. If you are not in a position to fulfill your obligations you should notify us immediately, otherwise we will be obliged to cancel our agreement.

We expect your answer as soon as possible.

Yours sincerely

**Figure 3. An example of the system output: *offensive tone*.**

the user has to insert a list of products by typing a string for each product, then this module produces the concatenation of these strings taking into account punctuation and conjunction aggregation. According to the user's selection about the tone of the letter, there may be many phrases that generate a certain component. This module selects one phrase from this set **randomly**. Therefore, it is possible different letters to be produced by the same input settings. Moreover, this module deals with **low-level linguistic phenomena** such as punctuation checking, conjunction aggregation between two phrases and capitalization of the first letter for each sentence.

## 5. Two examples of the system output

Figure 2 and figure 3 show two examples of the system output in Greek (there are also their translations in English) as well as the stylistic settings selected by the user for each one of them. In order to illustrate the stylistic variations of the produced letters, we chose the *complaints for the delay of an order delivery* thematic category. This category includes the following components according to its letter plan:

(a) Mandatory components

1. Notification of non delivery of the order
2. Request for immediate delivery of the order
3. Request for notification about a possible inability for delivering the order
4. Determination of the regular delivery date
5. Exhortation to reply

(b) Optional components

6. Request for contact
7. Threat of possible order cancellation
8. Notification of subsequent problems (i.e., due to the delay of the delivery)

The user is able to select which of the optional components will be included in the letter. In the examples

shown in figure 2 we chose to include components 6 and 7. Hence, the produced letters include the components 1, 2, 3, 4, 5, 6 and 7. The correct order of the above components is following:

| Section | Components |
|---|---|
| Cause: | 1-4 |
| Purpose: | 2-3-6-7-8 |
| Conclusions: | 5 |

Therefore, the final order of the produced letter will be 1-4-2-3-6-7-5 (i.e., omitting component 8), as it can be seen in the examples. Note that besides the choice for letter tone, the examples have been generated based on the same input settings by the user. Example 1 could be the first complaint letter for a delivery delay whereas example 2 could be the last letter of a series of similar letters that have not been replied.

## 6. Performance-conclusions

The presented system is fully-implemented. It has been developed in Visual Prolog and runs on a Pentium PC. Currently, it supports a user-interface in Greek and composition of business letters in Greek, as well as over 30 primitive thematic categories. The system produces letters in Rich Text Format encoding, that is a wide spread text formatting compatible with the vast majority of word-processors.

It has been tested on the basis of the black box evaluation technique, that is evaluating system output without looking inside to see how it works [1]. 15 people that write business letters on a daily basis have been used it for a week. About 90% of the produced letters was absolutely correct (i.e., without a single error). Moreover, all the users faced no problems on dealing with the system and considered the letter composition procedure whether fast or satisfying. On the other hand, it was observed that the end users considered the selection of a thematic category whether difficult or not suitable for their particular needs.

Regarding the quality of the produced letters, it has to be noted that despite the lack of pronouns and referring expressions, the generated text is pleasant to read and manages to give the sense of a human-composed text due to the careful selection of canned phrases and the creation of stylistic variations that differentiate the produced letters.

The main strength of the presented generator lies in the ability to handle stylistic variations on one hand and in the possibility of producing different letters from the same input settings on the other hand. As a more or less template system, it is able to produce letters by manipulating efficiently canned phrases and strings inserted by the user but it really doesn't know what is the meaning of this process.

As it has been mentioned previously, one of the most important activities in the NLG research is the construction of text plan libraries. We constructed such a library that includes the representation of several categories of business letters with the view of achieving as much abstraction as possible. Furthermore, the user is able to select what (s)he wishes to be included in the letter by guiding the system to generate short or long letters.

Additionally, our generator is based on a language-independent procedure and it can be ported easily to another language since we haven't used any intrinsic element of the Greek language. Actually, we believe that the implementation of an English version, that is our short-term goal, would face less problems than the Greek one, since the morphology of English is much more simpler than the corresponding Greek one.

Maintainability and expandability of our system is a very crucial point. In general, maintaining and updating template systems is hard, especially when the number of templates is too large. Nevertheless, as it has been already underlined, a concrete component of a letter may be common in different thematic categories. It is possible, therefore, to expand our system with a new category by using some already existing components and defining only a few new ones. Our short-term research focuses on simplifying the insertion of a new thematic category model as more as possible.

Finally, our long-term research will be the extension of our system by integrating work in style identification with the view of achieving automatic business letter categorization in terms of their written style and tone properties.

## Acknowledgement

This work was granted by the Hellenic General Secretariat of Research and Technology in cooperation with the Department of Electrical and Computer Engineering of the University of Patras under the contract EPET II-715. The views, opinions, and/or findings contained in this paper are those of the authors and should not be construed as an official HGSRT position, policy, or decision, unless so designated by other official documentation.

## References

[1] Allen J. Natural Language Understanding. 2nd Edition, The Benjamin/Cummings Pub. Co., 1995.

[2] Ashley A. *A Handbook of Commercial Correspondence*, Oxford University Press, 1992.

[3] Bateman J., E. Maier, E. Teich, & L. Wanner "Towards an Architecture for Situated Text Generation", in *Proceedings of the ICCICL*, Penang, Malaysia, 1991.

[4] Brettos T. & A. Agalianou-Brettou *Commercial Correspondence*, Eugenidis Pub., 1984, (in Greek).
[5] Coch J. & R. David "Representing Knowledge for Planning Multisentential Text", in *Proceedings of the 4th Conference on Applied Natural Language Processing*, 203-204, 1994.
[6] Goldberg E., N. Drieger & R. Kittredge "Using Natural Language Processing to Produce Weather Forecasts", *IEEE Expert*, 9(2), pp. 45-53, 1994.
[7] Hovy E. *Generating Natural Language under Pragmatic Constraints*, Ph.D. Thesis, Yale University, 1987.
[8] Hovy E. "Language Generation: Overview", in Cole R. *et al.* (eds.) *Survey of the State of the Art in Human Language Technology*, available in http://www.cse.ogi.edu/CSLU/HLTsurvey/, 1995.
[9] Kukich K. *Knowledge-Based Report Generation: A Knowledge-Engineering Approach*, Ph.D. Thesis, University of Pittsburg, 1983.
[10] Lang L. *Letter Writing: A practical guide to effective communication*, HarperCollins Publisher, 1994.
[11] Michos S., E. Stamatatos, N. Fakotakis & G. Kokkinakis "An Empirical Text Categorizing Computational Model Based on Stylistic Aspects", in *Proceedings of the 8th International Conference on Tools with Artificial Intelligence*, pp. 71-77, 1996.
[12] Michos S. & E. Stamatatos "*Dialogos-EPET Project: Requirements Analysis*", Technical Report No. 7.1.1, University of Patras, 1995, (in Greek).
[13] Reiter E., C. Mellish & J. Levine "Automatic Generation of Technical Documentation", *Applied Artificial Intelligence*, 9(3), pp. 259-287, 1995.
[14] Springer S., P. Buta & T. Wolf "Automatic Letter Composition for Customer Service", in R. Smith & C. Scott (eds.) *Innovative Applications of Artificial Intelligence 3*, pp. 67-83, 1991.