INTERNATIONAL JOURNAL OF TRANSLATION
Vol. 15, No. 1, Jan-June 2003

Robustness in the Linguistic Automaton

X.PIOTROWSKA,
R.PIOTROWSKI,
Y.ROMANOV
Herzen State Pedagogical University, Russia

INTRODUCTION

It has been argued that the 21st century marks a new era in the evolution of the homo sapiens who is going into a new period - the age of neoanthropos. This new human being has to live and function in cooperation with various types of informational technology and, first of all, with NLP systems. Thus, the computer has become an important instrument of communication and information processing.

Therefore the man-computer interaction has not only technical and language engineering aspects but also semiotic and even psycholinguistic and cognitive ones. The developers of the NLP and MT systems turned their efforts to working out some really functioning systems. In the process of working, they strove for creation of such technologies which would keep in better harmony with the laws of natural language. These ideas gave rise to behavior-based (BB) artificial intelligence and MT conceptions. New NLP and MT systems should be developed on the behavior-based principle of AI constructing and functioning which means that any NLP should be oriented to step-by-step elimination or, at least, reduction of ambiguity in linguistic solutions with the help of the aforesaid technologies. The user himself should also be involved in the process and, on the basis of his own semantic-pragmatic orientation, should actively help the system eliminate ambiguity. On these very principles, the international Speech Statistics group (SSG) has been developing a poly-functional multilanguage system "Linguistic Automaton" *(LINGTON).*

LINGTON is a thoroughly balanced complex of hardware, software and lingware means interacting with an extensive linguistic data- and knowledge base (KB). It is intended as a multi-purpose NLP system which should model, in a robust way, the verbal/mental behavior of humans in a particular social role: that of a translator, a text

interpreter and a language teacher. It should comply with the following requirements:

1) to be multifunctional, i.e. to be able to perform various kinds of text processing tasks, such as the initial statistical text processing, language recognition, spell-checking, indexing, annotation, abstracting, a man-machine dialogue and machine translation of oral and written texts;
2) to be able to minimize information losses when overcoming the language barrier between the natural language and that of the LINGTON;
3) to possess connecting devices to enable its links with other sources of information through communication channels, Internet, etc.;
4) to allow for further developments and improvements in its structure by adapting the LINGTON to the communication-informational evolution of society, to be adjustable to the changing pragmatic demands of actual users of information;
5) to possess robustness and vitality, i.e. a built-in ability to preserve its most essential properties in case of failure or malfunctioning of external devices, RAM breakdowns, distortion of words and text fragments, etc.

ROBUSTNESS AND VITALITY OF A LINGUISTIC AUTOMATON

Robustness and vitality of a linguistic automaton is considered as its ability to cope with the deficiency of information coming to its input which gives rise to deficient descriptions at some descriptive strata. This deficiency may be due either to some imperfection of the input itself or to shortcomings of the processing resources (e.g. disambiguation in phonetic/graphemic recognition, lexicon, grammar parsing, or semantic/pragmatic analyses). In this case, LINGTON may degrade the communication quality to the point of being competent to generate reliable results, as well as of being able to hold its own vitality.

Let us take up the robustness mechanism as applied by a linguistic automaton to producing a machine translation.

When creating a new robust linguistic automaton, the modular-hierarchical organization of NLP or MT of an oral or printed text is of prime importance. This technology provides for a step-by-step decrease of ambiguity coming from the initial and more primitive levels of processing to the higher levels of text analysis. At choosing the final solution, the higher levels have the higher priority than the lower levels

of analysis. The text processing procedure goes through the following levels:

1) phonetic/graphemic recognition,
2) the lexical-morphological (dictionary) level, where the analysis and, if necessary, translation is done of word forms (w/f) and of fixed lexicalized word combinations (w/c),
3) the micro-segmented level, where they carry out the analysis (and translation) of nominal w/cs and of verbal groups with the nucleus presented by a finite and non-finite forms of the verb,
4) the macro-segmented level, where the identification and processing of functional segments is reached (i.e. the subject, the predicate, the object and the adverbial groups),
5) the sentence level, where the syntactical structure of the input sentence is identified and the corresponding output structure is selected from the pre-loaded set of output frame structures,
6) the text level, where the final corrections and marking of the results of NLP or MT are carried out, proceeding from the analysis of the theme, structure and pragmatics of the text.

Thus, each block is responsible for recognition of the input linguistic units (LU) of the corresponding level, their descriptions and selection of the output elements. At that, on each level, the input sentence is transformed into a sequence of pairs of: input LUs *(u)*, i.e. w/fs or w/cs, + their lexical-grammatical or semantic/syntactical characteristics (χ) and output LUs *(u')* + their characteristics *(χ')*, i.e.

$$T = u_1\chi/u'_1\,\chi'_1,\ u_2,\chi_2/u'_2\chi'_2,\ldots u_i\chi_i/u'_i\chi'_i,\ \ldots\ u_n\chi_n/u'_n/u'_n\chi'_n\chi$$

On the second level, the above said LUs and phrase patterns receive their lexical-grammatical characteristics *% and %'* right from the automatic dictionary (AD). For units or groups of the higher levels, semantic and syntactic characteristics are passed from the lower levels or worked out in the corresponding block.

In the batch mode of NLP or MT of large flows of non-normalized and, sometimes, faulty texts, LINGTON constantly meets with "faulty" situations. They can be coped with either by the system itself or by way of the man-machine dialog. In the latter case, the full configuration of the MT block must ensure functioning of such operations as the following:

- inter-editing of and additions to the translation and, if necessary, its formation and re-formation. These operations are carried out with the help of the built-in editor, a portable scanner and a Quick-Link Pen,
- temporary stopping of the translating procedure, so that the user could look through the list of untranslated w/fs and w/cs, which the system did not find in its KB, and translate them himself and add them to the system's AD or thesaurus.

These methods of "socio-partnership" interaction of man and computer are directly tied to the task of maintaining robustness, vitality and self-training of the LINGTON. To solve the task, we have, first of all, to provide for the following:

- introducing regularly met with fragments of the input text together with their adequate and normative translations into the translator memory,
- adding to the set of frame patterns, ensuring syntactically and stylistically normative translations, which can also be used for didactic purposes,
- introducing additional probabilistic evaluations in the translation graphs of the LINGTON modular-hierarchical organization.

CONCLUSION

Let us take a look at the more typical cases of the system malfunctioning and the ways of coping with them for maintaining its robustness and vitality.

1. If several output schemes are received on level *n*, then:

- all the output variants are passed to level *n+1* in order to make ambiguity to be solved on the next level either by the user or by the system itself,
- the system selects that output variant which is structurally the closest to the semantic-syntactical scheme of the input and, because of that, requires a minimum of transformation.

2. If an NLP system suffers a fail in formation of the output text structure provided by its level *n*, then the user receives the results that the system worked out on the previous level. In other words, decision-making is based on the "synergetic" ability of the system to decompose

or simplifyingly modify the general task ***P***. In this case, the general task is presented as a set of separate sub-tasks:

$$\ddot{u} = (\ P_1, \ddot{u}_2, \ldots, P_i, \ldots, \ddot{u}_k).$$

As an example, let us take a look at the situation which we faced when developing an experimental Turkish-Russian MT system. Non-isomorphism of the nominal and the verbal Turkish and the corresponding Russian paradigms is very strong. So, the lexico-morphological modules were unable, without coordinated interaction with the modules of analysis and synthesis of the surface and deep structures of the sentence, to generate Russian w/fs and w/cs which would be morphologically correct and corresponding the input Turkish LUs. Since such modules for the Turkish-Russian MT were not created yet, we had to divide the lexical-morphological task *P* into three separate sub-tasks:

P_1 – the analysis of a Turkish LU, where it was divided into a stem (the initial form) and affixes which it was constituted of,
$\ddot{u}_2$ - identification of the grammatical nature of each affix,
P_3 - translation of the LU.

Turkish LU: EDINLEN

P_1 - *Structure*	$\ddot{u}_2$ - *Descriptor*	P_3 - *Translation*
EDIN-	the stem	(to) receive
-IL-	the passive voice indicator	
-EN	the participle indicator	
		received

Using information retrieved from the solution of each of the above-mentioned sub-tasks, the user himself formed the translation of the input Turkish sentence.

Another example of taking, by the system, an independent "synergetic" decision to change a general complex task *P* to its simplified modification $\ddot{u}^*$, is the MT system's transition to a word-for-word and phrase-for-phrase translation when it lacks morphological, syntactical and semantic resources for building the surface and the deep structures of the input sentence. Thus, decomposition and simplification of task *P* make the LINGTON robustness noticeably stronger allowing the MT system to find a way out of the deadlock situations when it is unable to process a text after the prescribed pattern.

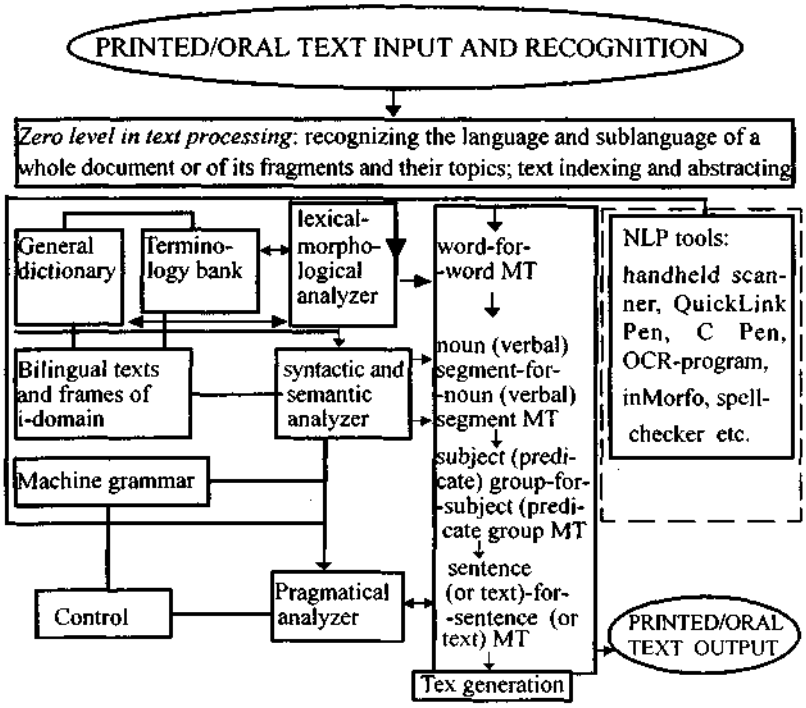

Fig. 1. Behavior-based model of NLP and MT

X.PIOTROWSKA,
R.PIOTROWSKI,
Y.ROMANOV
Herzen State Pedagogical University of Russia
48, Moyka Emb., St.Petersburg, 191186, Russia
E-mail: rp@yr4993.spb.edu