[*First International Conference on Language Resources & Evaluation*, Granada, Spain, 28-30 May 1998]

Russian Resources in Language Engineering : Evaluation and Description

Vera Semenova, ANALIT

Abstract

The paper presents the results of the survey of Russian resources in language engineering, the methods for language resources evaluation and description which were developed in the framework of the project. Evaluation procedures as well as the results are also considered. The evaluation criteria are systematically chosen from the end-user point of view.

SURVEY GENERALITIES

The survey of Russian resources in language engineering (LE) was carried out by Russian company ANALIT and French company SCIPER in 1994-1997, at the request of the French Ministry of Research. In June'94 the preliminary results were presented to the Russian LE-community (Semenova, 1995). The first European-wide presentation took place at the ELSNET Goes East &IMACS workshop in 1995 (Semenova et al., 1995).

The results of the survey were edited in a book form and distributed in France (Semenova & Fluhr, 1996). The first version of the book covered only the domain of written language processing whereas the second version (Semenova & Fluhr, 1997) is extended to the speech processing.

The book is written in French. It is entitled "Les Industries de la Langue dans les Pays de 1'ex-URSS : répertoire des acteurs et des produits". In publications in English the book is usually named "Catalogue (or General Directory) of Russian Teams and Products in LE".

SURVEY STRUCTURE

It was necessary, at the very beginning, to define the structure of the data collected. As the main goal of the work was to contribute to establish a co-operation of Russia with European countries, it was decided to gather, first of all, information on teams working in the LE-area. The book was considered to be a catalogue of Russians teams, that's to say a reference tool for French NLP-technologies producers willing to find Russian partners and Russian language resources (LRs).

It was decided to describe really existing teams, not in dependence of their formal affiliation, if any. In Western surveys only two types of teams are usually used: academic/industrial. But in Russia there exist many team which have the both "caps" or even more. Moreover, a lot of informal teams were found. For Western readers it seems important to realise that almost the half of teams listed in the Catalogue are informal. Even if they have formal affiliation in an academic institution, they consider their resources and products as their own property.

In order to show the background and the capabilities of each team it was supposed necessary to describe the teams' products. It must be underlined that this word was used in the survey in a very wide sense - it covers not only commercialised production but also prototypes, models etc. It becomes clear if remember that the goal of the work was not the sale of the software but the co-operation establishment.

Thus, the structure for the Catalogue was chosen as follows: two volumes, the first one consisting of team descriptions, the second one - of products descriptions. It was decided to divide the 2nd volume into chapters according to product type.

In the last version the 2nd volume contains the chapters corresponding to the following product types: information retrieval (IR) systems, terminology and dictionary management systems, spell-checkers, machine translation (MT) and computer-assisted translation systems, linguistic parsers, OCRs, speech synthesis systems, speech recognition systems and speech processing systems, oral dialogue systems, linguistic resources.

The two volumes are linked by cross-references. Each team description refers to the products, each product description refers to the producer.

INFORMATION GATHERING

In order to gather information a set of special questionnaires had been developed. The first of them is the questionnaire for team descriptions. Each type of product has its own questionnaire which takes into account the essential features of this type.

Method "interview" for filling-in the questionnaires in a dialogue form proved to be the more effective one, although it takes a lot of time and efforts. Distribution questionnaires by e-mail or by other telecommunications seems low effective; supplementary telephone contacts are always needed.

Now, when many teams have they own homepages on the Web, the information gathering became less difficult, but in the beginning of the project (1994-1995) we had no this opportunity.

The questionnaires have been originally written in French because they were developed in collaboration with French partners. But in Russia almost nobody understands French, so the questionnaires were translated into Russian and during all the information gathering period only the Russian texts of the questionnaires were used.

At first it was supposed that all catalogue texts obtained in such a way will be translated later literally into French. But the volume of the really gathered information seemed to be so great in comparison to the forecasted one that we had to invent a reduce format for the French catalogue. That's why the format of the French catalogue is significantly reduced in comparison to the Russian one: each description (a team's one or a product's one) occupies one page.

The questionnaires have been filled-in in Russian. The real filled-in questionnaires are often accompanied by other supplementary texts - in Russian, in English or (rarely) in other languages. People gave the texts they already had - scientific articles, program documentation, publicity, proposals etc.

In the last version of the Catalogue the 1st volume comprises the descriptions of 99 teams and also some indexes.

The 2nd volume contains the following numbers of product descriptions in the chapters :

Product type	**Number of products**
IR systems	20
Terminology and dictionary management systems	34
Spell-checkers	21
Machine translation and computer-assisted translation systems	16
Linguistic parsers	37
OCRs	16
Speech synthesis systems	7
Speech recognition systems and speech processing systems	13
Oral dialogue systems	2
Linguistic resources	

Table 1: Number of product descriptions in the Catalogue

It is difficult to say how many resources is presented now in the last chapter of the Catalogue, apparently more than 250. This chapter is organised not in the same way as other chapters where the description of each product occupies a page. In the chapter "Linguistic resources" all the resources of the same team are presented together.

LINGUISTIC PARSERS

The chapter "Linguistic parsers" seems to be the greatest and the most heterogeneous among the Catalogue chapters which are represented in the Table 1. It consists of a rather heterogeneous set of products. Obviously, the term "linguistic parsers" does not provide a strict definition of the bounds in which a product can **be** classified as belonging to this class because any system which accepts a text of any kind as its input and produces some representation of this text as its output after processing it with certain procedures including usage of linguistic models complies with the definition.

Strictly speaking, any linguistic computer technology has to contain elements of linguistic parsing, that is why we included in this chapter only those systems which were described by their authors as valuable by themselves, not taking into account whether they are incorporated in any larger system or not.

As the result of this rather loose definition, the following classes of linguistic systems can be found in this chapter

1. Systems of text indexing;
2. Systems of automatic compiling of thesauri and concordances;
3. Systems which form requests to databases on the basis of requests formulated in natural languages.
4. Systems of automatic analysis of semantics, syntax and morphology of natural language texts.

These classes don't cover all the products represented in the chapter. It seems desirable to make more detailed structure of product types within this chapter although it is rather difficult to give an exact classification because many linguistic parsers can be included in several classes simultaneously according to the practice and possibilities of their usage. However, here is a short characteristic of each of 4 classes above:

1. The first class is represented by products which provide the set of descriptors (keywords) on the basis of the input text and then index the text by these descriptors. There is a lot of various methods used to solve this problem. For example, there is a system which indexes texts without any morphological or syntax analysis, only by means of literal similarity.

There are also systems which use for the solving of similar problems morphological analysis which, due to the specificity of the problem, is usually restricted to singling out stems and subsequent lemmatisation.

2. The second class of systems is dedicated to processing large text corpora in order to create concordances and thesauri, used afterwards for text indexing and information retrieval.

It seems worth to point out here that, as it follows from indirect mentions, similar original programs for compiling wordlists (lexicons) by processing large text corpora have been developed by other teams as well. They were not described by the authors as separate products but considered by them as auxiliary tools for developing other linguistic technologies.

3. Systems of linguistic processing of natural language requests to databases are essentially a kind of machine translation with only one difference - instead of the natural language they use special machine language for requests to a database as their output. Therefore, the technology of creation of such systems is similar to the technology of MT development.

4. The last class of linguistic parsers consists of systems which realise one or several stages of text analysis. As a rule, such systems presuppose the possibility of their adaptation to definite tasks and their future incorporation in larger and more complicated systems: MT, information retrieval etc. Many of these parsers include original technological solutions and ideas.

In the chapter there are also syntax analysis systems, which produce syntactic representation of the input text. Practically all such systems use dependency tree for representation of syntactic structure of the Russian language.

Technologies for creation semantic representation of the text are also represented in the chapter.

Regarding the commercial usage of the products included into this chapter, it can be stated that most of them are designed not for commercial but for academic purposes. However there are systems which, although they are not sold as is, are incorporated into commercial products.

In the conclusion it can be said that programs and projects presented in this chapter though generally are not ready for immediate commercial usage, nevertheless can be of interest to those who are seeking for new technical solutions and technologies. Linguistic parsers solve more local problems then MT-constructing or information retrieval system development, and this allows their authors to spend more time on bringing to perfection their part of NL-processing. Thus it creates the possibility to introduce new quality to the technologies in which these parsers will be incorporated.

OCR EVALUATION

It seemed difficult to distinguish OCRs one from other using only their descriptions, as for this product type all the descriptions are similar. That's why it was necessary to find other criteria for their comparison and evaluation.

Here, as well as in other evaluations, we tried to find user-oriented criteria. That's to say, we wanted to find a way to evaluate linguistic technologies from the end user's point of view.

The OCR components the most important for the users are:

a) level of user interface accommodation;
b) "system-scanner" interface organisation;
c) cognitive technologies applied:
d) efficiency of recognition errors automatic post-corrections.

From our point of view, the component "c" is not only the most "intelligent" but also the most significant as it determines the potential possibilities of the system development. Moreover, the properties a), b), d) seem to go to be practically similar for all Russian OCR systems. So, our technology was oriented primarily to estimate the component "c") as the "recognising core" of the OCR system.

To be efficient in pattern recognition, OCR use simultaneously a whole complex of techniques:

=> to suppress "image noises";
=> to distinguish meaningful signs;
=> to use relations between meaningful signs for error correction.

We wanted to measure exceptionally the OCR ability to distinguish meaningful signs, independently of their forms (face and size of letter). So, we used various methods to eliminate the contribution of other factors in summary effectiveness of evaluated OCR For example, to eliminate context factor we compared the OCR tools by giving them for recognition several randomly constructed "pseudo-texts" consisting of the forms of the most frequently used Russian lexemes.

In our opinion, the common digital criterion (percentage of well-recognised symbols) usually used to evaluate the effectiveness of an OCR is not convenient from the user's point of view. Users are usually interested to know how many words have to be corrected after recognition. So, we proposed (Arapov et al, 1995) other criterion - percentage of well-recognised words, as the basic one of our technology.

It is evident that in the case of usual texts the value of our criterion can not be higher than that in the recognition of a "pseudo-text". Therefore, the pseudo-text compilation technique enables to predict error percentage.

But the optimal OCR is not obligatory that one having the minimal error percentage. In our opinion, the optimal OCR for industrial applications is the most stable OCR, i.e. the ideal one would be the OCR which reads the text properly and doesn't change the behaviour due to eventual font alterations.

Thus, our notion "OCR stability" includes, in particular, the following aspects which seem to be important to estimate an OCR::

1) size of the fonts' diapason where the OCR rests stable;
2) value of the above-mentioned criteria in the optimal area;
3) the mode of this value variations out of the optimal area (not only the number of errors but also the number of error types, etc.)

The experiments described in (Arapov et al, 1995) had shown that the following fonts' properties are essential to describe the optimal area:

=> monospace/proportionality;
=> wide/narrow letter holes;
=> size of symbols.

At the same time, for example, the presence or absence of serifs is less significant.

In our experiments, in order to minimise the number of pilot OCR evaluations we used sharp contrasted fonts' types.

In particular, it was found that the monospaced fonts lies in the optimal area of all the OCR systems in comparison.

Choosing the fonts for the experiments we took the fonts really used in Russia, as the OCRs seem to be the AI-systems the most closely connected with the properties of Russian language and national graphics as well as with national traditions of text printing and using of documents.

MT-SYSTEMS EVALUATION

The next attempt to find a method to compare and to evaluate NLP-technologies took place for the MT-systems.

Machine translation has long ago found its place in the market of linguistic systems. It is usually considered as a uniform problem, though the products which represent this type of linguistic technologies vary essentially both from the point of view of realisation of translation *per se* and from the point of view of their applications to certain fields and practical tasks. In the strict sense of the word, the term "machine translation" presupposes that the system, having received as an input a text in one language, produces in the process of its work the same text in another language without any human participation.

However, many years of development of such systems showed that it is hardly possible for the computer to produce translation in any degree close to the one produced by a human being. That is why almost all presently existing systems are oriented to perform a slightly different task, that is to construct such representation of the text in the end language of the user, that a person not acquainted with the original (for instance, not speaking the language in which the text is written) could get maximum information considering the meaning of the text in the minimal period of time.

With regard to this formulation of the problem, the main parameters of MT systems are as follows:

a) the quality of the translation, which is evaluated more in the terms of intelligibility of the text than by the grammatical correctness of the resulting translation.

b) The width of thematic fields and volumes of vocabulary supported by the system.

c) The number of supported languages.

d) Speed of the translation.

The most part of MT-systems, represented in the Catalogue, is built in the bounds of principles, formulated in the I. Melcuk's "Meaning " Text" model, which is based on the multilevel representation of language pins transition rules between these levels. Although it must be noticed that the degree of correspondence to this model varies from system to system.

Some of the systems use quite different language models, there are also systems built along the principles formulated long before the multilevel models

One can distinguish here commercial products, which try to combine quality, speed and prime cost of the translation, and thus often compelled to sacrifice the former of the parameters, and non-commercial, developers of which pay more attention to the quality of translation thus sacrificing speed and economy of their systems, which hampers their practical usage in a great degree.

From the point of view of the languages processed by Russian machine translation, English is the undoubted leader, it is supported by practically all systems.

The thematic fields, covered by MT-systems are concentrated around the fields which have maximum demand. One of them is constituted by scientific/technical texts of various domains, the other are business correspondence and mass-media. The field of technical translation is the most developed one. The switch between different genres is usually managed by connecting dictionaries for the domains.

If consider the translation quality as the criterion of the evaluation, one can see that this notion is rather indefinite and equivocal. Usually it means the adequacy of the senses and styles of the both texts (the source one and the target one), under the condition that the translation has been done without grammatical errors. But everybody who seen at least once the texts produced by MT-systems, understands that it is useless to try to find grammatical and stylistic adequacy.

The stylistic adequacy can be neglected at the first phase of the study, because the texts in electronic form are mostly used for information transmission, therefore the most important are the grammatical correctness of the translation and its semantic adequacy to the source text.

Actually there is no MT-system capable to translate without grammatical errors. Therefore, the grammatical correctness of the translation can be used as the criterion, the more so the sentences, translated correctly, in most of cases have the same sense as the text source. The main difficulty in dealing with this criteria is to estimate the gravity of grammatical errors which torture the contents. One can suppose that the gravity of the same errors

varies in dependence of the type of the text, so the problem becomes more difficult.

But the choice of criterion depends on the user's needs. Grammatical correctness and adequacy could be criteria for the written translation. Such kind of translation usually needs post-editing, therefore the time spent for it could serve as a criterion.

For the communications and information transmission the most important criterion is certainly the comprehensibility of the translation. This criterion correlates very well with the adequacy, because it is low probable to have alternative understanding on a rather long part of text.

So, this criterion was chosen for our evaluation tests which were organised as follows. Two MT-systems, Stylus and Socrat, were evaluated in comparison. Some test texts had been translated by the both MT-systems, and the translations were given to 3 users. Each user had to mark comprehensible sentences by "+" and incomprehensible by "-" Then the pluses and the minuses were counted.

The tables 2 and 3 present the results of the evaluation by one of the users.

	STYLUS		**SOCRAT**	
	+	-	+	-
test 2	10	0	8	2
test 3	6	4	5	5
test 4	15	6	12	9
test 5	56	11	51	16
test 6	41	11	39	13
TOTAL	128	32	115	45
TOTAL %	80 %		71,9 %	

Tabl.2. Test results: numbers of comprehensible (+) and non- comprehensible (-) sentences for user S.

	STYLUS	**SOCRAT**
test 2	100	80
test 3	80	60
test 4	71	54
test 5	83	76
test 6	79	75
totals:	82,6	69

Tabl.3. The percentage of comprehensible sentences evaluated by user S.

STYLUS **SOCRAT**

	+	-	+	-
UserS.	56	11	51	16
UserN.	57	10	49	18
UserV.	52	15	51	16

Tabl.4. Comparison of evaluations made by different users (for the test 5)

One can suppose that the evaluation results vary significantly when done by several users. The table 4 presents the results of comparison of evaluations made by different users, for one of the texts.

Certainly, in order to make serious conclusions, much more tests have to be carried out, with more participants. Our goal was to suggest an appropriate evaluation criterion based on the user's needs.

REFERENCES

Arapov, M.V. et al. (1995). How to Compare and to Evaluate OCR Systems? (our Approach). In Proceedings of the ELSNET Goes East and IMACS Workshop on Integration of Language and Speech, November 9-11 (pp. 5-10). Moscow, Russia: Institute for Information Transmission Problems RAS.

Semenova, V. (1995). Linguistic technologies in Russia. Science & Technology in Russia, 2(8), 31 - 32.

Semenova, V. (1998). Russian Resources in Language Engineering. ELRA Newsletter, 2 (3). To be published.

Semenova, V. & Fluhr C. (1996). Les Industries de la Langue dans les Pays de 1'ex-URSS: Répertoire des Acteurs et des Produits. Version 1. Paris, France: MESR (DISTNB).

Semenova, V. & Fluhr C. (1997). Les Industries de la Langue dans les Pays de 1'ex-URSS: Répertoire des Acteurs et des Produits. Version 2. Paris, France: SCIPER.

Semenova V. et al. (1995). Survey on Russian Teams and Products in Language Engineering: Impressions, Facts, Conclusions. In Proceedings of the ELSNET Goes East and IMACS Workshop on Integration of Language and Speech, November 9-11, Moscow (pp. 164—168). Moscow, Russia: Institute for Information Transmission Problems RAS.