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

# Some Problems in the Evaluation of the Russian-German Machine Translation System MIROSLAV

**Jutta Marx & Nancy Smith & Bernhard Staudinger**
FG-Informationswissenschaft
Universität Regensburg
93040 Regensburg, GERMANY
[*firstname.lastname*@sprachlit.uni-regensburg.cle]

## Abstract

In this article some of the problems encountered in designing an evaluation for the MIROSLAV machine translation (MT) system will be examined. We will concentrate mainly on the strategies employed in the evaluation and the evaluation areas, namely translation quality and translation productivity, whereas the former will be treated in more detail than the latter. In conclusion we will present two new methods for measuring translation quality through intelligibility and informativity. For the former we propose a novel sort of performance task, for the latter an automatic comparison of the semantic features assigned during the machine translation process to source and target texts. Both methods should reduce the subjectivity of judgments of translation quality. The former, because it measures intelligibility indirectly, that is, test persons are not required to make a direct judgment and the latter, because it will compare information units (semantic features) automatically.

## 0. Introduction

MIROSLAV (Machine Translation Initiative Russian and other Slavic Languages) is a joint research project funded by the German Ministry for Education and Research (BMBF). The broader goal of the project is to introduce the standards and methods of computerized language processing to the languages of Eastern Europe. The main practical aim of the project is to develop a Machine Translation System for Russian and German (and eventually other Slavic languages). Partners in the project include: GMS/Berlin; SNI/Munich (Siemens-Nixdorf); the Slavic Department at Humboldt University, Berlin and the Information Science Department (IWR) at Regensburg University. The main tasks of the IWR are lexical-semantic solutions for complex transfers and the evaluation of the Russian-German system. Evaluation partner will be GESIS, Berlin. Project reports (in German) from the IWR group can be found at *http://rsls8.sprachlit.uni-regensburg.de/-smn22575/miropage.html*. The current status of the MIROSLAV project is that a Unix-based Russian-German prototype has been developed and a PC-based demo version was presented at CeBIT 1998. The project is scheduled to run until July, 1999.

In this article some of the problems encountered in designing an evaluation for the MIROSLAV machine translation (MT) system will be examined. The evaluation is still in the planning stage, although some concrete preliminary work has been completed, including a test evaluation and a first examination of the workflow of the evaluation partner.[1] Due to space limitations not all aspects of the evaluation can be presented in detail here. We will concentrate mainly on the strategies employed in the evaluation and the evaluation areas, namely translation quality and translation productivity, whereas the former will be treated in more detail than the latter. In conclusion we will present two new methods for measuring translation quality through intelligibility and informativity. For the former we propose a sort of performance task, for the latter an automatic comparison of the semantic features assigned during the machine translation process to source and target texts. Both methods should reduce the subjectivity of judgments of translation quality. The former, because it measures intelligibility indirectly, that is, test persons are not required to make a judgment and the latter, because it will compare information units (semantic features) automatically.

## 1.0 MT Evaluation Design: Some Strategies and Considerations

### 1.1 Evaluation Strategies

The literature distinguishes various evaluation strategies or approaches. The different strategies often have points in common, although they generally do vary with regard to their objectives. In the current article we will concentrate on those strategies relevant to the MIROSLAV evaluation design. What is presented here is not meant to be an exhaustive list of MT-evaluation strategies.[2]

- The MIROSLAV evaluation will be **individual** (versus comparative), that is, we will be testing the MIROSLAV system alone, not in comparison with other MT systems.
- The MIROSLAV evaluation will be mainly **functional** (versus **formal**)[3], due to its dual aims of determining translation quality and productivity. But it will also include formal factors, such as an error classification, for determining the linguistic quality of the output. A **functional** approach takes into consideration the area of application of the MT program, since the value of a system can only be seen in relation to this.[4] The following aspects of a **functional** MT evaluation can be distinguished:

    1. For whom will the evaluation be performed? In the case of MIROSLAV, the evaluation will be **user-oriented**.
    2. What is being evaluated?
       The **complete system** within its computing environment (as opposed to its individual parts: lexicon, translation core, etc.)
    3. What aspects of the system are being evaluated?

[1] cf. Marx (1998) for more information about the preliminary evaluation work completed in the MIROSLAV project so far.

[2] cf. Staudinger (1998) for a State of the Art of Evaluation approaches.

[3] In the sense of Vasconcellos (1988).

[4] Formulated for the first time by Bar-Hillel (1971). The first functional MT evaluation was attempted by Henisz-Dostert (1978).

As mentioned above, our evaluation will concentrate on **translation quality** and **translation productivity.**

4. Which evaluation method will be employed? The MIROSLAV evaluation will employ a **global, black box** (versus glass box) evaluation method[5] with access to the lexicon, but not to the translation core, nor to the analysis or generation modules. Further, the evaluation will be both **direct** and **indirect[6].**

- The MIROSLAV evaluation will be a **macro** evaluation[7], the aim of which is to determine the adequacy of the system output within its environment - without diagnostic statements. This is a classic example of a black box approach.
- The MIROSLAV evaluation will be an **operational** evaluation in the sense of Way (1994), who distinguishes three different strategies: **typological, declarative** and **operational.** In the operational evaluation (also **economic[8]** evaluation), parameters that do not necessarily concern the MT system itself are of importance. The integration of an MT system into the workflow of a specific user, the so-called *set-up*[9] of a translation process, is of great importance here. MIROSLAV is of the operational type in that we intend to measure the translation productivity of the system by examining the workflow of our evaluation partner.

### 1.1.1 Evaluation Typology

To sum up the MIROSLAV evaluation strategy we turn to an evaluation typology proposed by Bourbeau (1990). We have added some points to Bourbeau's typology to characterize the MIROSLAV evaluation. An explanation of the individual levels of the hierarchy follows the figure.

- Determining the *user's requirements* is at the top of the hierarchy. For Bourbeau, this mainly means the desired translation quality. He distinguishes two classes of translations: *informational translations*, which are only used internally, and *translations for further use*, i.e.

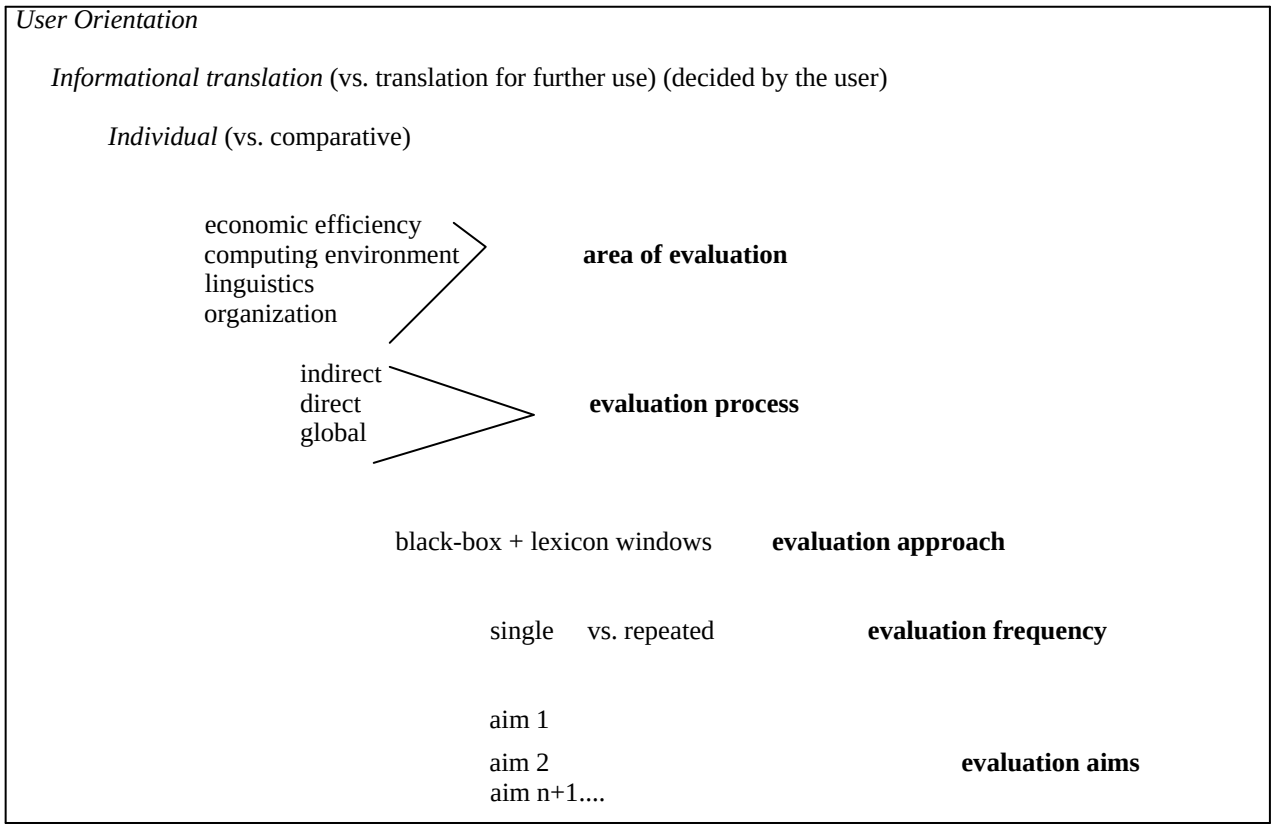


Figure 1: Typology of the MIROSLAV evaluation

[5] For further discussion of the method or design question for MT evaluations see King (1995)
[6] In the sense of Bourbeau (1990).
[7] In the sense of Van Slype (1979).
[8] cf. Galliers & Sparck-Jones (1996:76f), who include, in addition to the *operational* evaluation, an *economic* evaluation. Cf. also Van Slype 1982, who proposes an analogous methodology under the term "Evaluation de l'efficacité" (evaluation of effectiveness)
[9] cf. Galliers & Sparck-Jones (1996:11) and Brey (1997:11).

translations which will be published. The MIROSLAV evaluation intends to concentrate mostly on informational translations.

- The next level in the hierarchy distinguishes between *individual* and *comparative* translations. This is a point not covered in the Bourbeau typology. MIROSLAV will evaluate individually.
- Bourbeau then distinguishes four focal points according to the area of evaluation: *economic efficiency* (increase in productivity through the use of MT), *computing environment* (compatibility and integratability within the user's existing computing system), *linguistics* (translation quality and linguistic performance), *organization* (integration into the workflow, necessary restructuring, etc.). The MIROSLAV evaluation will cover all four of Bourbeau's areas of evaluation: through evaluation of translation productivity, we will investigate the areas of economic efficiency, computer science, and organization; the evaluation of translation quality will cover the linguistic area.
- For the *evaluation process* Bourbeau distinguishes three phases: one *indirect,* which aims at gaining knowledge about the system by questioning users, through software manuals, etc.; the second or *direct* phase has access to all relevant areas of evaluation; the third or *global* phase incorporates all insights gathered in the previous two phases. The MIROSLAV evaluation will incorporate all three phases of Bourbeau's evaluation process.
- The *evaluation approach* distinguishes between *glass-box* and *black-box* evaluations. As mentioned above, glass-box evaluations allow the developers full access to the software and all other components of the MT system. Black-box evaluations are carried out by the user, who usually has only limited information about the software itself, but can draw conclusions about the practical use of a translation system from the test results. The MIROSLAV evaluation will be of the black box type with access to the lexicon.
- In the next level of the hierarchy, the *frequency of evaluations* is determined: is a single evaluation desired, or should the improvability of the system be established through successive upgrades and evaluations? For determining the translation quality, MIROSLAV will employ a single evaluation. Translation productivity, on the other hand, will be determined with successive evaluations and upgrades.
- As a last point Bourbeau looks at the *evaluation aims.* The aim of the MIROSLAV evaluation is to determine to what extent the MT system in question can be profitably integrated into the workflow of the evaluation partner.

## 1.2 The areas of evaluation

### 1.2.1 Criteria for translation quality

There are many criteria for judging the quality of machine translation. This section will give an overview of some of **those** criteria and their measures as they apply to the MIROSLAV evaluation. We judge translation quality according to the rough translation output text of the system, that **is,** only the German target text will be evaluated. Here are some of the criteria for judging the quality of machine translation:

- **Informativity (accuracy, fidelity):** This criterion deals with the question of how much of the information content of a text is lost during the translation, or, positively formulated, how much information is kept. Many diverse methods for measuring information have been proposed. Each involves a comparison of the source and target texts. The most well-known is a 10-step descriptive scale developed by Carroll (1966) within the framework of the ALPAC evaluation. The individual points on this scale and general conditions will not be discussed in detail here.[10] The greatest problem with this rating method is the subjectivity inherent in both the definition by the evaluators of the individual points on the scale and in the judgment of the test persons. Some attempts have been made to reduce the subjectivity by statistic means, using a very large test population. Another method to combat subjectivity in the rating method is a limitation of the scale and a precise definition of the individual points. In the MIROSLAV evaluation the "usability"[11] of the translated texts is defined by the needs of the evaluation partner. Thus, as mentioned above, we will be concentrating on informative translations, which do not lend themselves to the informativity criterion. One exception is the planned automatic comparison of semantic features in source and target texts. (see section 2.2.1.3)
- **Intelligibility:** The most common technique for measuring intelligibility is, again, a rating method. Carroll (1966), for example, employs a nine-point scale.[12] The problems mentioned above (precision, subjectivity) also hold true in this case. We propose a simple three point scale to measure intelligibility, which should limit the subjectivity of the judgments somewhat. The multiple choice questionnaires of Leavitt et al. (1971) are not suitable for the MIROSLAV evaluation situation. Direct questioning of the test person, the so-called *comprehension*[13] or *knowledge*[14] *test* can also be used to measure intelligibility of a translation. Tests of this type measure not only translation quality, but also the intelligence of the test person. The *performance test,* in which test persons are asked to carry out the described actions, tests not only translation quality and the intelligence of the test person, but also his or her aptitude. Despite this drawback a type of performance test will be employed in the MIROSLAV evaluation in the form of the assignment of key words to the translated texts. The subjectivity of the test persons will be reduced in that they will not be required to make any direct judgment of informativity. (see section 2.2.1.1)
- **Error analysis of the output:** Error analysis is not an evaluation criterion in the same sense as, for example, intelligibility. It is more of a measure of the overall quality of a translation. It is based on the rather simple consideration that a relation between quality and the

[10] For Carroll's (1966) scale cf also Falkedal (1991) or Van Slype (1979).

[11] in the sense of Lenders(1978) "Nützlichkeit".

[12] Different scales for measuring intelligibility have been proposed, among others, by Van Slype (1977) and Leavitt et al. (1971).

[13] cf. Henisz-Dostert 1973 and Van Slype (1979: 79f). Jordan (1994) offers an example of how such a user questioning could look like.

[14] cf. Sinaiko and Klare (1972,1973).

number of errors exists. One problem with error analysis is the exact definition of an error. Another difficulty is that errors may be of differing quality (e.g. stylistic vs. lexical): thus merely counting errors is not sufficient. Error rankings or classifications must be called into play. Most error classifications are linguistically motivated. The MIROSLAV evaluation will use a modified version of Flanagan's (1994) error classification for error analysis.

- **Post-editing:** For post-editing, the same is true as for error analysis: it is not an evaluation criterion in the strictest sense, but rather a measure for the overall quality of a translation. The basic consideration is that a relation exists between the time needed for post-editing and the quality of a translation. The amount of time and energy needed to make a machine translation qualitatively indistinguishable from a human translation is measured. Different methods have been proposed to measure this amount, e.g. counting the number of necessary corrections per page, measuring the amount of time needed to correct a page, counting the number of machine-translated words remaining after correction, or counting the number of keystrokes needed to correct the machine output.[15] The post-editing criterion may be implemented in a later stage of the MIROSLAV evaluation if requested by the evaluation partner.

#### 1.2.2 Criteria for Translation Productivity

The productivity of a translation system depends on many factors: on the one hand the translation software itself, on the other hand the system environment, but also on the translated text and the user. Therefore criteria for translation productivity can be divided into three areas: those concerning the hardware environment (RAM, CPU, etc.), those having to do with the translation software itself (ergonomics, etc.) and user-specific criteria, which involve, among other things, the workflow in which the MT system will be integrated. In the MIROSLAV evaluation we will concentrate mostly on user-specific criteria and software ergonomics. In this paper only the former criterion will be covered.

##### 1.2.2.1 User-specific criteria affecting translation productivity

- **Workflow:** The way in which a translation system is integrated into the workflow of the user or company using MT has a strong influence on the productivity of a system. This also poses the question of possible restructuring within a company (e.g. new work areas for pre- and post-editing, the division of labor, etc.).
- **Text type:** Productivity (and quality) of an MT system are in direct relation to the text type to be translated. Texts with many unknowns may require updating the lexicon, which takes time. De- and reformatting of graphs, illustrations, or charts will also take time.
- The evaluation of productivity must also take into consideration the **system** and **hardware knowledge** of the user. These factors may have an effect in regard to trouble shooting or required training. Methods to determine these factors are user interviews and user tests.
- **user motivation:** factors of a psychological nature may influence the productivity of a translation process. Employees attitudes towards the introduction of MT may effect productivity. Pre- and post-editing of machine output may require repeated correction of identical errors and quickly result in frustration. These aspects are difficult to test and should be sorted out through exhaustive questioning and consultation before evaluation begins. The extent to which these aspects need to be considered for a concrete evaluation is not entirely clear.[16] In the MIROSLAV evaluation preliminary results of user-motivation have revealed a very positive reaction to the introduction of MT.

### 1.3 Other Considerations

#### 1.3.1 Evaluator Interests

The vested interests of evaluators are of great importance in determining the entire design of an evaluation and may lead to varying, often incompatible results. As for MT, there are certain special qualities which contrast with the ISO 9126[17] standard for software products. At least six interest groups with inevitably varying evaluation interests can be identified: sponsors, manufacturers, system developers, competitors, researchers and users[18]. As mentioned above, the MIROSLAV evaluation is exclusively user-oriented. For this reason only the special user interests will be reviewed here.

##### 1.3.1.1 User-oriented evaluation: defining the user group

The ISO norm 9126 puts the user at the center of the evaluation. The current transfer of commercial MT systems from workstation environments to PCs has resulted in the problem of the inhomogeneous nature of the user group. While the user group used to be limited to professional translators and specialists in the workstation environment, since the transfer of many MT systems to a PC environment the user group can no longer be so easily defined. In regard to translation quality and productivity, it is crucial for every evaluation to define an exact user profile since quality and productivity demands may vary greatly among different users. For this approach evaluators must bear in mind that generalizations about the quality and productivity of an MT system are not possible. User evaluation interests can be manifold, ranging from the general employment of MT up to selection criteria for a specific system.

#### 1.3.2 Test Material

Adequate data, commonly in the form of texts, is needed to evaluate MT systems. The main questions in this regard are what kind of texts should be used, how much material is adequate and which text types should come into consideration. These parameters can have a strong effect on the results of an evaluation.[20] Contrary perhaps to expectations, it is quite difficult to gain access to machine-read-

[15] cf. Church & Hovy (1993:243) and Su et al. (1992).

[16] cf. Hess (1997).

[17] cf. Staudinger(1997).

[18] cf. Staudinger (1997) for a detailed discussion of the evaluation interests of these various groups.

[19] cf. also Steiner 1993, who discusses an evaluation differentiated according to interest groups.

[20] cf. Falkedal (1991:22ff).

able Russian texts. For this reason we decided to create a domain or user-specific corpus for the evaluation, partially with the aid of a scanner and OCR software. This will be a collection of specific texts which will not lay claim to representativeness for the Russian language, but will nevertheless be representative of the texts used by the evaluation partner. The effort inherent in creating such a corpus should not be underestimated. Collecting the texts, either through scanning or other means, and transliteration of the Cyrillic alphabet have caused unforeseen problems and taken much time.

#### 1.3.3 Test Persons

Some remarks about the test persons are due at this point. In MT evaluation there is a preference for so-called *independent experts* as test persons. The term independent expert refers to persons without MT experience or even translation experience, in general. The idea is that such people will approach the evaluation in a more objective way. This touches on one of the general problems with test persons in an evaluation: objectivity. Another decisive question concerns the number of test persons. In order to gain the highest possible objectivity, it is considered necessary to have a large number of test persons. The choice of these test persons should also be representative. Another factor that comes into play is the test persons' knowledge of languages. For the judgment of informativity, where input and output texts must be compared, test persons must have command of both languages involved in the translation. For judgment of readability, on the other hand, test persons should be monolingual.[21]

## 2.0 The MIROSLAV Evaluation

### 2.1 The evaluation partner

#### 2.1.1 Choice of the Area of Application

The prerequisite for the choice of an evaluation partner was work with Russian language texts. Dealing with the largest possible amount of texts from different areas was also considered desirable in order to have access to a wide range of texts. Concerning productivity, growth potential should be present, i.e. new work areas might be made possible by a reduced human translation workload. Thus not only would the number of texts processed increase, but also the effectiveness of the complete system. As an evaluation partner for the MIROSLAV project we were able to win the Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. in Berlin, in particular the department for "Information Transfer in Eastern Europe" (GESIS Berlin). This institute turned out to be extremely suitable for the evaluation purpose because of an existing text variance and the area of application of the texts. A description of the institute and its main work areas follows.

#### 2.1.2 Description of the Evaluation partner

##### 2.1.2.1 Objectives and main tasks of GESIS Berlin

Since 1992 GESIS has had a branch in Berlin, which is mainly concerned with keeping track of the results of social science research in the former GDR. They also deal with the transfer of information on social science topics to and from Eastern Europe. Furthermore, they give advice on questions concerning social science research methodology and on ways of gaining information in the new federal states and in Eastern Europe.

Within the framework of information exchange between Eastern and Western Europe, GESIS Berlin regularly carries out inquiries about institutes and projects in Eastern Europe. The acquired data, along with information gleaned from printed materials (e.g. journals), are published and distributed (e.g. in the newsletter "Sozialwissenschaften in Osteuropa"). Part of this information is stored in a database on social science research projects (FORIS) at the Informationszentrum Sozialwissenschaften in Bonn (IZ Bonn).[22]

#### 2.1.3 Description of the materials used and those produced by the evaluation partner

##### 2.1.3.1 Potential texts for machine translation

Numerous questionnaires in the Russian language from a survey, carried out every second year, concerning past, planned, or current research projects in Eastern Europe are evaluated by GESIS. These FORIS questionnaires are filled out by researchers, either by hand or typewriter. GESIS Berlin regularly reviews journals, essays, and monographs in the Russian language with the goal of finding information concerning conference proceedings, references and research projects, institute descriptions, etc. This material is available in good printed quality (not machine readable).

Texts from the Internet, e.g. the homepages of Eastern European institutes, are employed to create profiles of institutes, which appear in the newsletter "Sozialwissenschaften in Osteuropa" or are stored in the IZ database. Relevant net addresses acquired in this way are entered in a clearinghouse (in English) for the social sciences, section on Eastern Europe, run by GESIS.

##### 2.1.3.2 Potential uses for information gleaned from machine translated texts

GESIS Berlin publishes a number of printed materials and maintains several electronic databases on topics of interest to the social sciences in the broadest sense. Furthermore, they enter and maintain the data acquired during their activities in several databases. The materials and databases most relevant to MIROSLAV are described briefly below. The newsletter "Sozialwissenschaften in Osteuropa" (The Social Sciences in Eastern Europe) is published at least four times a year, in German and in English. It contains current information on the social sciences in Eastern Europe, for example[23]:

- social science institutes and the focal points of their research
- profiles and indexes of relevant journals on the social sciences (reviews of some relevant articles)

[21] For a more elaborate discussion of these problems see Falkedal (1991:20ff) or Van Slype (1979:166).

[22] For more information on GESIS Berlin see Marx/Mutschke/ Schommler (1995) or the WWW homepage http://www.berlin.iz-soz.de/

[23] source: http://www.berlin.iz-soz.de/publications/newsletter/socsci-eastern-europe/index.htm

- short contributions on specific questions of the social sciences in Eastern Europe
- references to monographs or studies, researchers, databases
- information on scientific societies and current conferences

The project database FORIS (Forschungsinformationssystem Sozialwissenschaften) is maintained by the IZ Bonn and contains information about past, current, and planned research on social science topics. Apart from social science projects in Eastern Europe, the database covers mainly the German-speaking area. The inventory presently consists of about 35.000 documents.

### 2.2 The design of the MIROSLAV evaluation

After finding an evaluation partner with the guidelines outlined above, we set out to custom design an MT evaluation for this user. The user-orientation of the evaluation determined several aspects of the design. The evaluation partner is mainly interested in the informative quality of roughly translated texts (without postediting) and in how MT might boost translation productivity within the department. As mentioned above, the material to be translated was also determined by the user. The use to which the test material (texts in the Russian language) is put by the evaluation partner, e.g. the assignment of key words for entries in the FORIS database, opened up several unique possibilities for testing translation quality. The evaluation design, as mentioned above, is still in the planning stages. Most of the planning for evaluating translation quality has been completed and will be the topic of the rest if this paper.

#### 2.2.1 Evaluation of Translation Quality

The evaluation of the translation quality of the test material from GESIS, can be divided into three main parts according the criterion for evaluation: intelligibility, error classification and informativity.

##### 2.2.1.1 Measuring the intelligibility of the translation

A group of naive test persons will evaluate the intelligibility of the roughly translated texts, sentence by sentence on a three point scale. The judgments will *be fully intelligible, partially intelligible, unintelligible.* The simplicity of the scale should simplify the task of judging and reduce the subjectivity of the test persons' judgments. The second method of testing the intelligibility involves the assignment by the independent experts of key words to the texts. This is a sort of performance test, as explained in section 1.2.1. The list of key words, or descriptors, will be those used to classify texts in the FORIS data base. The independent experts will be GESIS workers, who are familiar with the key word assignment task, but not with the translation of texts.

##### 2.2.1.2 Evaluation by means of error classification

The MIROSLAV evaluation team will perform two main tasks to judge translation quality. Firstly, a sentence by sentence error analysis will be carried out, with the aid of a language pair-specific error classification based on Flanagan (1994) and Rinsche (1993). As in other parts of the evaluation, the error classification will be designed from the point of view of the user and will only deal with errors, which negatively affect intelligibility, that is, not questions of style or word order, insofar as they are neutral in respect to intelligibility.

##### 2.2.1.3 Measuring the informativity of the translation

Secondly, to test the informativity of the translation an automatic comparison of the semantic features assigned by the MIROSLAV MT system to the source and target texts will be undertaken. This point demands a bit more explanation. The MIROSLAV MT system employs semantic features in several phases of the translation. In the lexicon of the MIROSLAV MT system, semantic features are assigned to the following four categories of words: nouns (TYN), verbs (TYV), adjectives (TYA) and prepositions (TYPREP).[24] In the analysis phase of the translation, combinations of these features may lead to new features being assigned to positions in the phrase structure tree, e.g. the combination of the TYPREP and TYN of a prepositional phrase will always result in a PTYPE, or prepositional phrase type. The system assigns values to these features as the analysis proceeds. We propose to automatically compare, sentence by sentence, the semantic features assigned in the source text to those in the target text. The idea behind this comparison is that the semantic features of source and target texts reflect the information content of those texts. This is, as far as we know, a completely new method of measuring the informativity of a translation. One must, of course, take into account that this evaluation method is system dependent, that is, it cannot be universally applied. We also recognize the limitations inherent when no one-to-one relationship between the two languages exists, for example, a concept may be expressed by a verbal phrase in the source language and by a deverbalised noun in the target language. Despite these limitations we believe this method will provide significant results.

## 3.0 Conclusion and Outlook

At present there is no single universal evaluation method for MT systems. The variety and heterogeneity of existing MT systems in regard to their conception, architecture, and above all their areas of application pose the question of whether a single, universal method is either possible or desirable. In general it could be said that the purpose of an MT system, namely to provide a reasonable translation of a text, should constitute a relative standard by which the quality of the translation output could be defined. The exact measure of this quality is, however, controversial and varies among interest groups. Both translation quality and translation productivity are relational not absolute values because they are user-dependent.[25] We have presented a user-oriented evaluation method which reduces the subjectivity of translation quality judgments by using two new methods, firstly through the use of a novel sort of performance test and secondly, through the automatic comparison of system assigned semantic features.

As mentioned above, work on the evaluation design is continuing. Several questions about the evaluation design remain to be clarified, among others the exact number of

[24] TYN=Type of Noun, TYV=Type of Verb, TYA=Type of Adjective, TYPREP=Type of Preposition.

[25] For further discussion of this relation see Staudinger (1997).

test persons and the amount of test material to be employed. These points and the measurement of translation productivity will be the subject of the next phase of the evaluation part of the MIROSLAV project, scheduled to be completed by July, 1998.

## Acknowledgments

The work described in this paper is part of the joint research project MIROSLAV, funded since mid-1995 by the BMBF (Bundesministerium für Bildung, Forschung, Wissenschaft und Technik). The Universität Regensburg has the project number 01 IN 508 E2.

## References

ALPAC. (1966). Pierce, J.R., Carroll, J.B., et al. *Languages and Machines. Computers in Translation and Linguistics. Report of the Automatic Language Processing Committee, Division of Behavioral Sciences. National Academy of Sciences.* National Research Council. Publication 1416. Washington, D.C.

Bar-Hillel, Y. (1971). Some Reflections on the Present Outlook for High Quality Machine Translation. In: Lehmann, W. P. & Stachowitz, R. (Eds.), *Feasibility Study of Fully Automatic High Quality Translation.* Austin: LRC.

Bourbeau, L. (1990). *Elaboration et mise au point d'une méthodologie d'évaluation linguistique de systèmes de traduction assistée par ordinateur.* Rapport final. Secrétariat d'État du Canada.

Brey, T. (1997). *Design eines Tools zur Verwaltung und Evaluierung von Testsuiten.* MA-Arbeit, Regensburg.

Carroll, J. B. (1966). *An Experiment in Evaluating the Quality of Translations.* Washington, DC. National Academy of Sciences. ALPAC report, Appendices 10 and 11, (pp. 67-78).

Church, K.W. & Hovy, E.H. (1993). Good Applications for Crummy Machine Translation. *Machine Translation* 8, 239-258.

Falkedal, K. (1991). *Evaluation Methods for Machine Translation Systems - An Historical Overview and a Critical Account.* Draft Report, ISSCO, Genf.

Flanagan, M. (1994). Error Classification for MT Evaluation. AMTA 94.

Galliers, J.R & Sparck-Jones, K. (1996). *Evaluating Natural Language Processing Systems.* Berlin et al.: Springer.

Henisz-Dostert, B. (1978). User's Evaluation of Machine Translation: Georgetown MT System, 1963- 1973. Paper presented at the Workshop on Evaluation of Machine Translation Systems. Luxembourg.

Henisz-Dostert, B. (1973). Users' Evaluation of Machine Translation. Rome Air Development Center, RADC-TR-73-239. New York.

Hess, M. (1997). *Einsatzbedingungen von Maschineller Übersetzung im Unternehmen.* Universität Zurich, Institut für Informatik.

ISO/IEC 9126. Information Technology - Software Product Evaluation-Quality - Characteristics and Guidelines for their Use. December (1991).

Jordan, P.W. (1994). A First-Pass Approach for Evaluating Machine Translation Systems. In: Falkedal, K. (Ed.), *Proceedings of the Evaluators' Forum, 1991, Les Rasses, Vaud, Switzerland.* Genf: ISSCO. (pp. 1-23).

King, M. & Falkedal, K. (1990). Using test suites in evaluation of MT systems. In: Association for Computational Linguistics. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics.* Vol. 2. (pp. 211-216). Pittsburgh, PA.

King, M. (1995). The Evaluation of Natural Language Processing Systems. In: *Proceedings of Swan 21, Geneva.* (pp. 97-109).

King, M. (1996). Evaluating Natural Language Processing Systems. *Communications of the ACM* 36. 73 —79.

Leavitt, A.W.; Gates, J.L. & Shannon, S.C. (1971). Machine Translation Quality and Production Process Evaluation. RADC-Technical Report (pp. 71-206).

Lehrberger, J. & Bourbeau, L. (1988). *Machine translation: linguistic characteristics of MT systems and general methodology of evaluation.* Amsterdam: Benjamins.

Lenders, W. (1978). Bewertungskriterien für maschinelle Sprachübersetzungssysteme. CEC, Luxembourg.

Marx, J. (1998). *Vorarbeiten zur Evaluierung der Übersetzungsqualität und -produktivität von METAL (russisch-deutsch).* In: MIROSLAV/R-Bericht 5/3. Regensburg.

Marx, J., Mutschke, P., & Schommler, M. (1995). *Möglichkeiten der intelligenten Integration heterogener Datenbestände - Das Projekt GESINE.* Bonn: Informationszentrum Sozialwissenschaften der ASI.

Rinsche, A. (1993). *Evaluationsverfahren für maschinelle Übersetzungssysteme: zur Methodik und experimentellen Praxis.* Technical Report, Kommission der Europäischen Gemeinschaften, Bericht EUR 14766 DE.

Sinaiko, H.W. & Klare, G.R. (1972). Further experiments in language translation: readability of computer translations. *ITL* 15. 1-29.

Sinaiko, H.W. & Klare, G.R. (1973). Further experiments in language translation: a second evaluation of the readability of computer translations. *ITL* 19. 29-52.

Staudinger, B. (1998). *Evaluierung von MT-Systemen: A State of the An.* In: MIROSLAV/R-Bericht 5/3. Regensburg.

Staudinger, B. (1997). *Zur Evaluierung von MT-Systemen: grundsätzliche Überlegungen.* In: MIROSLAV/R-Bericht 3/3. Regensburg.

Steiner, E. (1993). Producers - Users - Customers: Towards a Differentiated Evaluation Research in Machine Translation. *Machine Translation 1.* 281-284.

Su, Keh-Yih, Wu, Ming-Wen & Chang, Jing-Shin (1992). A New Quantitative Quality Measure for Machine Translation Systems. *Proceedings of COLING-92, Nantes.*

Van Slype, G. (1979). *Critical study of methods for evaluating the quality of machine translation. Final report.* Bruxelles: Bureau Marcel van Dijk.

Van Slype, G. (1982). Conception d'une méthodologie générale d'évaluation de la traduction automatique. *Multilingua* 1(4). 221-237.

Vasconcellos, M. (1988). Factors in the Evaluation of MT: Formal vs. Functional Approaches. In: Vasconcellos,

M. (Ed.), *Technology as Translation Strategy*. Binghampton: State University of New York. (pp. 203-213).

Volk, M. (1997). *Evaluierungskriterien für Übersetzungssysteme*. Universität Zürich, Institut für Informatik.

Way, A. (1994). Developer-Oriented Evaluation of MT Systems. In Falkedal, K. (Ed.), *Proceedings of the Evaluators' Forum, 1991, Les Rasses, Vaud, Switzerland*. Genf: ISSCO. (pp. 237-244).