

# End-to-End Evaluation of Machine Interpretation Systems: A Graphical Evaluation Tool

Susanne J. Jekat, Lorenzo Tessoro

University of Hamburg  
Computer Science Department<sup>1</sup>  
Vogt-Kölln Str., 30  
e-mail: jekat,lorenzo@nats.informatik.uni-hamburg.de

## Abstract

VERBMOBIL as a long-term project of the Federal Ministry of Education, Science, Research and Technology aims at developing a mobile translation system for spontaneous speech. The source-language input consists of human speech (English, German or Japanese), the translation (bidirectional English-German and Japanese-German) and target-language output is effected by the VERBMOBIL system. As to the innovative character of the project new methods for end-to-end evaluation had to be developed by a subproject which has been established especially for this purpose. In this paper we present criteria for the evaluation of speech-to-speech translation systems and a tool for judging the translation quality which is called Graphical Evaluation Tool (GET)<sup>2</sup>.

## Introduction

The performance of evaluation very often is driven by the characteristics of the system that has to be judged (Andenfilger, 1994). As to the Verbmobil project the evaluation should meet three aspects:

- the needs of the developers,
- the needs of the user,
- the constraints on the evaluation of translation quality in general.

In our concept and performance of evaluation we tried to combine these three aspects but one should keep in mind that at least the constraints on translation quality in general were meant to describe human translation with all its varieties and specific stylistic features. As to the special case of machine interpretation still only texts from limited domains can be transferred. So in our view it seems to be legitimate to simplify some of the procedures that are applied to the evaluation of human translation. An evaluation method based on any well known standard (EAGLES, 1995; Spark Jones and Galliers, 1996; Manzi, 1996) could not have integrated the three cited aspects, as traditional evaluation methods are intended for comparative evaluations more than for the investigation of a system during its development; therefore, to meet the requirements we had, we developed an integrated methodology and a tool for speech-to-speech quality evaluation which also allows easy access to the data.

## Translation Quality

Evaluation of translation quality is a complicated matter per se. One reason may be the diversification of the field, another the different manifestations of translation (from

simultaneous interpretation to written translations). In this paper we will refer to dialogue interpretation, that is the transfer of spoken language 1 to spoken language 2. In our view, interpretation differs clearly from written translation:

1. The process of writing a translation can take as much time as necessary (within a reasonable time-frame), but interpretation has to be fast because the ongoing communication should not be disturbed.
2. Written translation has to be very clear in order to avoid misunderstandings. The receiver of a written translation in contrast to the receiver of an interpretation has no contact to the translator nor access to situational or pragmatic cues to solve ambiguity. On the other hand, written translation offers the opportunity to use footnotes and explanations, which for the condition of time pressure can not be used in interpretation.
3. The input for the translation of written texts is complete and well-formed. Deviations from the well-known and well documented standard of written texts in most of the cases are no mistakes but motivated by special intentions of the author. Different from this, the input for interpretation is incomplete and sometimes ill-formed as compared to the structures of written language. Additionally, specific characteristics of a speaker or the actual situation may influence the appearance of the input.

For these reasons, the claim that translation should preserve form and function of the original text is only applicable to written translations. An imagined 'perfect translation' should contain every information of the source-language text. A deviation from this constraint should only be triggered by differences between source and target languages themselves (e.g. black grapes in English have to be referred to as blue grapes in German).

---

<sup>1</sup> This work was funded by the Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01 IV 101 A/O and in the framework of the SFB 538 Mehrsprachigkeit (Collaborative Research Center No. 538 Multilingualism) by the Deutsche Forschungsgemeinschaft (DFG). The responsibility for the contents of this study lies with the authors.

<sup>2</sup> To simplify the presentation of this paper, we only refer to the language pair German-English.

We are still far from an objective measurement for a 'perfect translation' and perhaps this goal cannot be reached because form and function are not independent (e.g. maintaining of the source-language function in the target-language text can interfere with the generation of an optimal target-language form) and researchers and human beings in general make different judgements on the importance of form or function.

As far as speech-to-speech interpretation is concerned, in our view the preservation of the text function is more important although function is not a fixed notion but has to be adapted to the concrete purpose of the actual communication. Speech-to-speech translation appears to be a classic case of covert translation where the texts to be translated are designed for consumption rather than edification and changes to form and content are possible in the interest of maintaining the function of the source-language text (House, 2000).

We outlined some principal constraints on the evaluation of translation quality and some characteristics of interpretation. The special needs of a user of a machine interpretation system are discussed in the following section.

### User-Oriented Evaluation of Translation Quality

VERBMOBIL is dedicated to facilitate communication between speakers of different mother tongues by generating an adequate translation.

In a dialogue, a speaker has two important tasks:

- to receive all important messages from the other speaker,
- to reach the communicative goal of the conversation.

As mentioned above, translation in speech-to-speech communication should support this by generating a target-language text which preserves all important messages and communicative functions of the source-language turn and which is enabling the ongoing conversation. Except in some rare cases, a word for word translation is no adequate strategy here.

### Input and Output Quality

Above the criterion of simply continuing the communication and beyond the struggle for a perfect translation, the user of a machine interpretation system should feel comfortable with the linguistic features of the system's output. In machine interpretation, where no pragmatic cues can be processed, the output is dependent on the input quality, so input and output in our evaluation are analysed following three linguistic criteria:

1. syntactical correctness of input/output,
2. semantical correctness of input/output,
3. possibility of misunderstanding of input/output.

The design of this evaluation phase ignores the translation relation, input and output have to be judged separately even though we suppose that there exists a relation between input and output (i.e. translation) quality. By this procedure we try to avoid circularity of analysis.

### Quality of Machine Interpretation

As to the translation quality of our system, the first step of analysis is the detection of a possible translation mismatch which consists of a loss or change of information in the translation process. This judgement is effected by a simple yes/no decision which is verified by a second step of analysing the translation quality. In this second step, information elements of the input are compared to those of the output from the user's point of view:

1. count of all information elements in the input,
2. count of essential information elements in the input,
3. count of information elements lost during the translation process (comparison between number of all input elements and number of those that are preserved in the output),
4. count of all information elements in the output,
5. count of additional information elements in the output as compared to the input.

For example, the turn '*when will we meet?*' in our domain (cooperative negotiation dialogues in the travel planning domain) consists of five information elements:

1. **when**: wh-question referring to time,
2. **will**: tense marker,
3. **we**: both speakers will do s.th. together,
4. **meet**: central goal of the communication,
5. **dialogue act (request-suggest)**: the whole question motivates the hearer to suggest a time for the meeting.

The different steps of evaluation are put together to a complete analysis of input, output and translation quality which has to be confirmed by a final judgement:

- a) a translation should be judged as '**Good**' if it is correct and if it does not contain any mismatches,
- b) a translation should be judged as '**Intermediate**' if it contains mistakes or mismatches but communication is successful,
- c) a translation should be judged as '**Bad**' if it contains mistakes and/or mismatches and communication is interrupted.

### Developer's Needs

End-to-end evaluation has to focus on the interaction between user and system in order to deliver reliable results of the system's performance under realistic application conditions. Besides the comparison of different systems, the central goal of any evaluation is to enable developers to improve parts of or the whole system. As there do not exist systems with a comparable architecture to VERBMOBIL, our evaluation focuses on the latter effect. In the following, we will describe those features of VERBMOBIL that determine the structure of evaluation.

### Characteristics of the System

The VERBMOBIL system is completely speech driven and can be used either in face-to-face communication or via telephone. The spoken input is processed by different speech recognizers (which themselves are compared in a separate acoustic evaluation) and then proceeded to several different models of translation used in

VERBMOBIL. The output always consists of only one of the translations, which is selected by the actual system's configuration. In the framework of this paper it would lead too far to describe all translation models. We only present two examples here:

1. the dialog-act based translation: the generation of the target-language turn is based on the central function (the dialog-act) of the input and some domain specific information elements,
2. the translation is based on a deep analysis of every lexical item and the syntactic structure of the source-language turn.

As evaluators should judge from the user's point of view and should not be influenced by a preference for one of the translation models, we test different configurations of the system but show them only one result at a time. Examples for configurations are *first wins* (the first result that is generated is taken into consideration) and *waiting for deep*, where the result of the deep analysis gets a very high priority.

During one evaluation phase (which normally lasts two weeks) the different configurations are tested in comparable numbers of test suites.

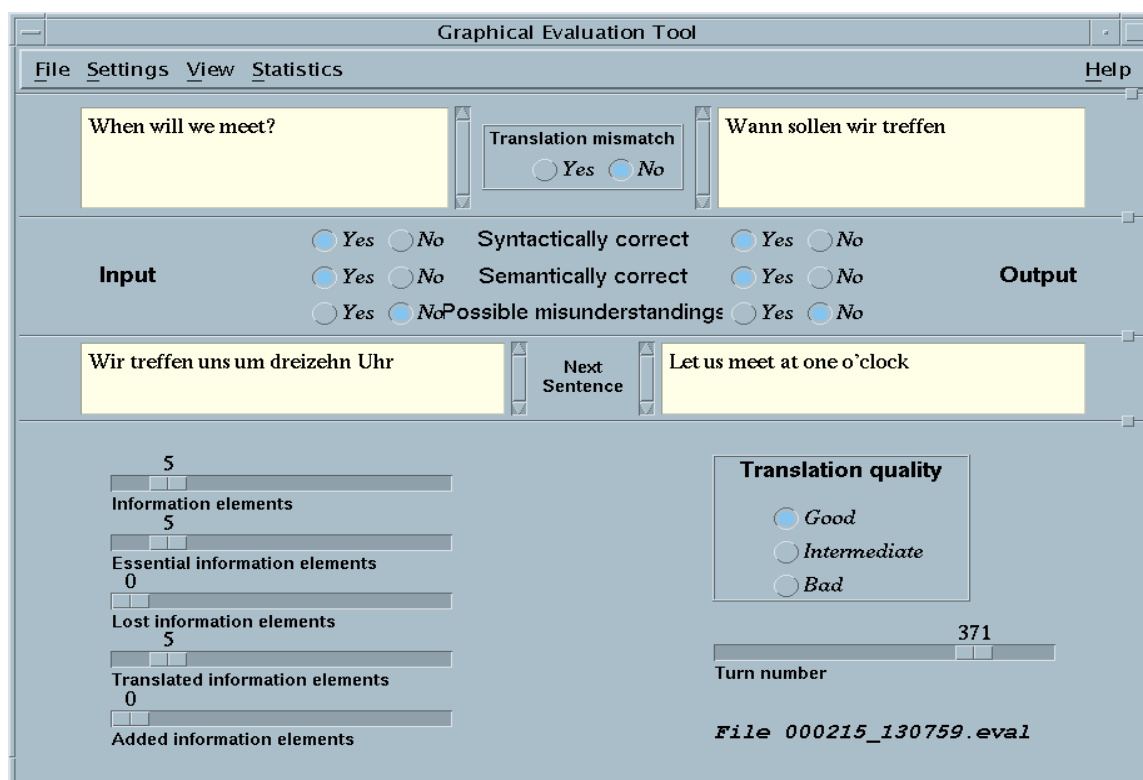


Figure 1: Main window of the GET

### Feedback to Developers during Evaluation

Within one evaluation phase more than one version of the system is tested as changes or updates are possible. The test results are continuously transmitted to the developers and they are informed immediately if there arise problems with a certain configuration. During the tests, evaluators and developers try to find out the weak points of a module or a module combination; changes are implemented and installed whenever possible throughout the evaluation phase.

### Feedback to Developers after Evaluation

Once an end-to-end evaluation phase is finished, all results are analysed under different aspects and presented to the

project members through the Graphical Evaluation Tool GET (see figure 1 above); the use of the same tool both for the evaluation procedure and the presentation of the results allows complete transparency of the evaluation method and a strong effectiveness. Ongoing improvement of the system is then based on the latest evaluation results.

### The Design of Experiments and the Evaluation Tool GET

During the end-to-end evaluation the VERBMOBIL system is tested by English and German native speakers, and the translations of the system are judged on the basis of the criteria mentioned in the section 'Translation Quality' (see above). Evaluators use the Graphical

Evaluation Tool, *GET*, which facilitates judgements according to these criteria (see also Jekat et al., 1999).

### Methodology of Experiments

Each test session consists of a dialogue between an English and a German speaker. The communicative goal of the conversation is to plan a joint business trip. Both speakers do not need to know the other language, VERBMOBIL functions as machine interpreter.

As visual cues are not processed by the system and therefore should not influence the ongoing communication, subjects are isolated in different rooms and do not see each other. They can only communicate by the help of the system. Both speakers are equipped with time tables and hotel lists, for every topic to be discussed they indicate their impression of having reached an agreement or not on a special sheet. A supervisor attends the tests, he does not interact with the speakers and writes a protocol of every test session. The dialogues are recorded and transcribed. The transcriptions and the log files of the system serve as input for the evaluation tool GET. Bilingual speakers or human interpreters then judge all relevant linguistic aspects of the dialogues.

### The Graphical Evaluation Tool (GET)

The end-to-end evaluation of bilingual dialogues that are interpreted by the VERBMOBIL system is performed by the help of the GET which is specially developed for this purpose.

Four different turns are displayed at the same time. The first represents the input from one of the dialogue partners (in our example the American English speaker), the box to the right of it displays the German translation of VERBMOBIL which is actually analysed.

The two boxes below show how the conversation is going on, the answer of the German speaker in the left and the English VERBMOBIL translation in the right box. By the help of the GET the translation is then evaluated according to all criteria described in the section 'Quality of Machine Interpretation' (see above).

As the tool is almost completely mouse-driven, it is very easy to use and it facilitates the processing of a large database (during the second phase of the VERBMOBIL project more than 300 dialogues had to be evaluated). The possibility to display a sequence of four turns at the same time allows the evaluators to consider also the dynamic features of the dialogues and to get an idea of the ongoing communication as a process.

Some separate windows of the GET show the main internal interfaces of the system.

The first window shows all the translations produced by the different translation modules of the system and highlights the segments which compose the final output. Evaluators can decide whether the chosen segments are the best option or not.

The second window shows the output of the speech recognizer for the current turn and highlights the wrongly recognized words. The tool can produce statistics for speech recognition providing the word accuracy, the

complete list of the used words for both languages, the occurrence and recognition rate for each uttered word and the sentence context for the words which were not recognized.

Two further windows provide the support for statistical computation: one window allows the user to put constraints on the statistics computation by choosing a pattern that a turn should match in order to be considered in the analysis; the other window shows the statistical results. For example it is possible to see which is the percentage of 'good' translations based on syntactically incorrect input or to measure the percentage of 'intermediate' translations for which there is a translation mismatch and which are syntactically incorrect. This procedure allows the investigation of the relations between different evaluation criteria.

### Results

With the help of the GET many dialogues could be analysed and relevant linguistic characteristics of the translation quality have been isolated. We also got information on the relation between the criteria related to the features of the system and the criteria related to the quality perceived by the users; we present here the results related to this topic.

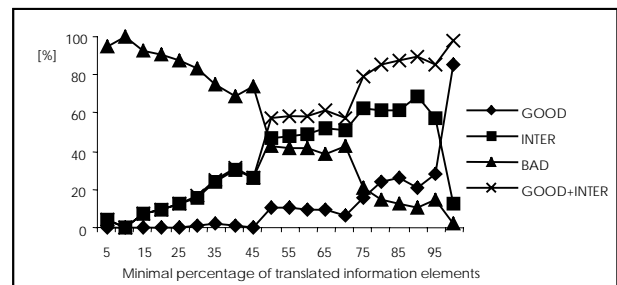


Figure 2: Effect of translated information elements on user oriented judgement

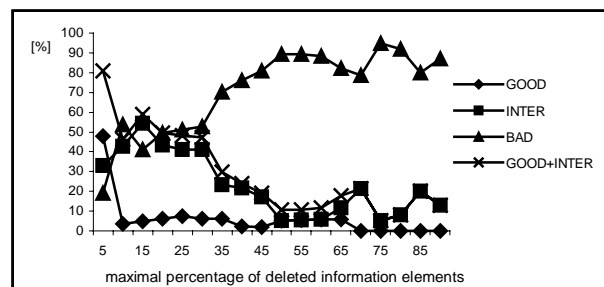


Figure 3: Effect of insertion of information elements on user oriented judgement

In figure 2 the relation between the quality of translation and the number of translated information elements is shown. The same relation is shown in figure 3 concerning inserted information elements. The value axis shows the percentage of turns with different translation qualities; the

category axis shows the minimal percentage of translated information elements and the maximal percentage of deleted information elements in one turn for figure 2 and 3 respectively. The graphics show that translation quality is much more sensible to the insertion of information elements than to the deletion.

Figure 2 demonstrates that when at least 50% of information elements are preserved in the translation the rate of 'good' or 'intermediate' translations is greater than the rate of 'bad' translations. On the contrary figure 3 shows that an insertion of more than 20% of information elements results in the predominance of negative judgements on translation quality.

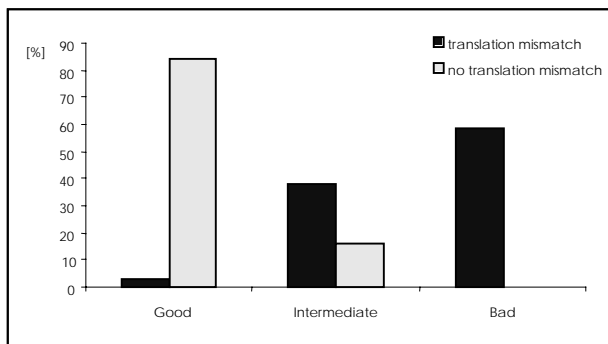


Figure 4: Effect of translation mismatch on user oriented judgement

Figure 4 shows the relation between translation quality and the occurrence of translation mismatches. Some translations which contain mismatches are even judged as good translations whereas in general, the absence of a translation mismatch seems to be the central criterion for a judgement as good translation. The translation mismatch is a sufficient criteria for a negative translation quality but it is not a necessary one: 40% of turns have an intermediate translation quality although they contain a translation mismatch. This means that these translations still help continuing the conversation.

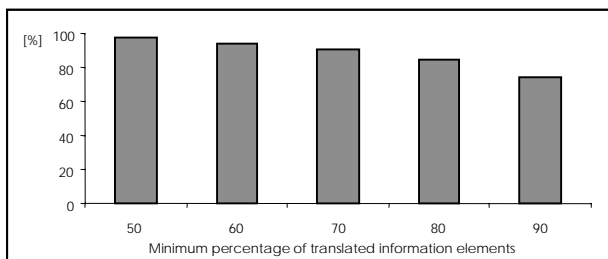


Figure 5: Cumulative distribution of translation mismatch for the minimum percentage of translated information elements

Figure 5 presents the relation between translation mismatch and a loss of information elements. The category axis represents the minimal percentage of translated information elements; the value axis represents the percentage of turns that contain a translation mismatch.

The more information elements are lost the more translations mismatches occur. Even if this is obvious, it is interesting to notice that the loss of a very low number of information elements can easily cause a translation mismatch.

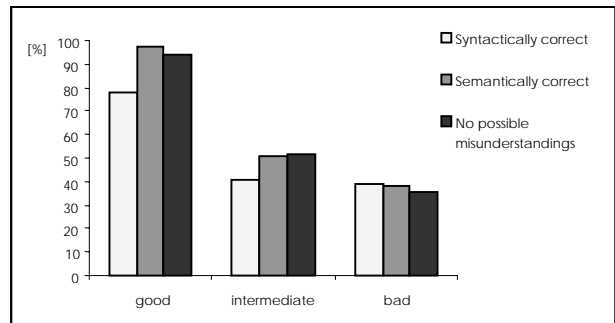


Figure 6: The quality of output for different translation qualities

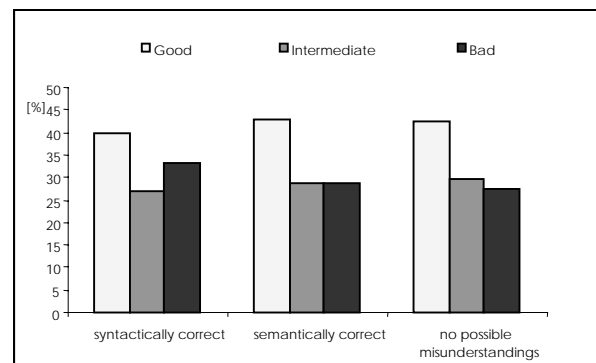


Figure 7: Translation quality for positive output quality

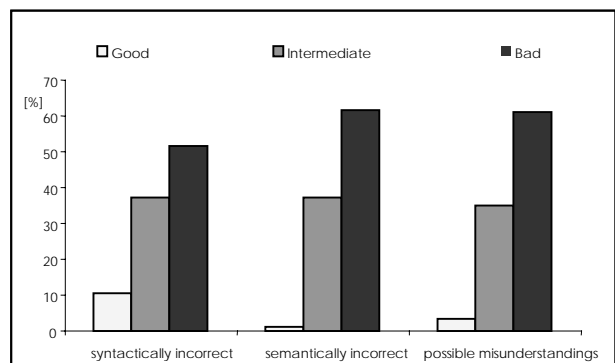


Figure 8: Translation quality for negative output quality

Figure 6 shows the quality of the output related to the translation quality. The value axis shows the percentage of turns with different translation qualities and linguistic judgements. There exists no significant correlation between semantic correctness and the absence of possible misunderstandings and a high translation quality but we suppose that the former are more important than syntactical correctness. This becomes evident in figure 7

and 8, where the value axis shows the percentage of turns for different translation and output qualities.

The correlation between quality of the output, translation mismatch and translation quality is shown in figure 9, where all the analysed turns contain a translation mismatch.

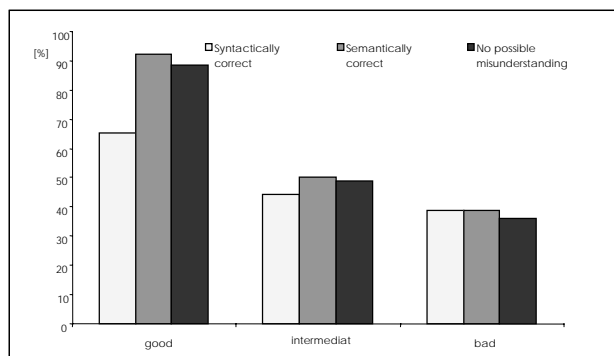


Figure 9: Distribution of output quality when a translation mismatch occurs for different translation qualities

## Conclusions

The GET tool allows a visual representation of dialogues that is useful both in the evaluation phase and for the improvement of the system because a large database can be processed. As to translation quality of machine interpretation, our evaluation method reveals that the preservation of the ongoing communication as a qualitative criterion is more important than a possible loss of information caused by a translation mismatch. When the output quality is related to the translation quality, semantic correctness is preferred to syntactic correctness. Despite individual differences between human evaluators all of them follow the same course of evaluation by the help of the GET. Statistical analysis of the results as well as quantity of the evaluated turns and number of different evaluators in our view lead to a relatively objective evaluation of translation quality performed by VERBMOBIL.

## References

- Andelfinger U. (1994). Some Remarks about the Validation of Information Systems Development. In Interdisciplinary Foundation of Systems Design and Evaluation. Seminar Report 97 of International Conference and Research Center for Computer Science, edited by R. Keil-Slawik, Saarbrücken Schloss Dagstuhl, University of Saarland.
- EAGLES-Expert Advisory Group on Language Engineering Standards (1995). Evaluation of Natural Language Processing Systems, Final Report. Document EAG-EWG-PR.2. Obtainable from ISSCO, University of Geneva.
- House, J. (2000). Übersetzungsäquivalenz: ein Schlüsselbegriff in der Übersetzungswissenschaft? In Knapp, K. and Knapp-Potthoff, A. (eds.) Sprachmitteln und interkulturelle Kommunikation. München: iudicium.
- International Standard ISO/IEC 9126. (1991). Information Technology - Software product evaluation - Quality characteristics and guidelines for their use. International Organization for Standardization, International Electrotechnical Commission.
- Jekat, S.J., Tessiore, L. and Lause, B. (1999). Das Graphical Evaluation Tool für die End-to-End-Evaluation des Verbmobil Systems. Verbmobil-Techdoc. Nr. 73, Universität Hamburg.
- King, M. (1996). Evaluating Natural Language Processing Systems. Special edition of Communications of the ACM on Natural Language Processing, 39(1), pp. 73-79
- Lehman, M.M., (1980). Life Cycles, and Laws of Software Evolution. In Proceedings IEEE, 68(9), pp. 1060—1076.
- Manzi, S. King, M. and Douglas, S. (1996). Working towards user-oriented Evaluation. In Proceedings of the International Conference NLP+IA/TAL+AI, Mouncton, Canada (pp. 155—160).
- Sparck Jones, K. and Galliers, J.R. (1996). Evaluating Natural Language Processing Systems. Berlin, Springer-Verlag.