[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland; ed. Kirsten Falkedal (Geneva: ISSCO).]

Some practical experience with the use of test suites for the evaluation of SYSTRAN

Ulrich Heid and Elke Hildenbrand
Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung – Computerlinguistik

1 Introduction

The purpose of this paper is twofold.[1] On the one hand, we try to give an overview of some results of practical work on the evaluation of the linguistic performance of SYSTRAN's French → German translation, concentrating on our use of test suites for the purpose of the partial evaluation of this system. Secondly, we try to discuss from a more general point of view the usefulness of test suites for this exercise and for different types of evaluation activities in general.

Much time and effort has to be spent on the construction of test suites, and therefore the question has to be addressed under which conditions and in which situations test suites can be most efficiently used as a linguistic evaluation tool, and where other, perhaps less complex, tools may lead to useful results as well.

This short paper starts with a general overview of major types of evaluation situations, concentrating on those types which mainly aim at the description of the *linguistic* part of a machine translation system.

We then describe our own work on SYSTRAN and discuss some of our experience with a test suite for verbal subcategorization and for certain other grammatical phenomena, such as tense, embedded structures, etc.

[1] This paper describes some results of a project carried out in cooperation with the headquarters of the German Federal Railway Corporation, Deutsche Bundesbahn, Frankfurt, the system developers of the SYSTRAN versions held at the Commission of the European Community, Telindus S.A., Luxembourg, and, in the initial project phase, the GMD, Gesellschaft für Mathematik und Datenverarbeitung. The project has been carried out in several phases, since 1988. We would like to thank Armin Schmidt (Frankfurt) for numerous suggestions and discussions. All misconceptions and errors remaining in this paper are of course our own.

We finally interpret theses results in terms of the usefulness of test suites for the different types of evaluation situations introduced at the beginning of the paper. We argue that test suites are a useful tool for cyclic long term evaluation as an accompanying measure which runs in parallel with system development; our results suggest, however, that it is less likely that test suites will be widely used for "snapshot evaluation", i.e. in evaluation situations where the goal is to come up with a rough idea of the performance of a given system at a given point in time.

2 Evaluation scenarios

Evaluation exercises are mostly performed on behalf of a “client”. Different types of clients may have diverging needs in terms of topics to be addressed in the evaluation. Consequently, the results may have to come more in a quantified form or more in the form of a general qualitative picture, depending on the client.

Evaluation clients

In terms of “clients”, we can distinguish a number of possible actors ranging from those involved most directly in practical use (“end users”) or in design work with a machine translation system to those supervising or sponsoring research and development on this topic.

“End users” of the output of machine translation may eg. be translators working regularly with a given system or deciders in translation and documentation companies responsible for acquiring new tools for their collaborators. They will be interested in the performance of a given system at a given time, or in a comparative overview of the performance of several systems. Their interest will focus less on the general linguistic performance of the system(s) (e.g. the precise description of all of the fragment covered, the linguistic pertinence of intermediate representations, etc.), than on how well the system’s performance meets their needs, e.g. in terms of sublanguage syntax and lexicon. Most end-users will be interested in a snapshot view of a system, i.e. of its performance at a given point in time.

The same evaluation needs are expressed by end users who do not themselves produce “linguistic services” but are consumers of such services. For example, our partners in the evaluation work described in this paper are in part interested in the immediate availability of “raw translations” for information purposes, even if the linguistic quality of the translation output is lower than one expects from human translation. The purpose of these raw translations is to allow deciders and technical specialists to assess the need for a full human translation of a piece of textual information. The decision about the need for a full translation is based, here, on the amounts of new and relevant information found in the texts. The MT system providing this service must produce lexically and terminologically adequate output, but the capacity to treat large syntactic fragments correctly is not the most important aspect, here. In terms of evaluation needs, the assessment of the quality of the lexical and terminological resources of a system, their adaptability, extensibility and use within the overall system is most relevant.

System developers, less directly involved in the use, but more in the *production and upgrading* of systems, are interested in additional topics: for developers, it is not only the performance at a given point in time that matters, but also, more difficult to assess, possibilities of enhancing and extending the system. In terms of linguistic performance, this means that system developers wish to know the starting-point for further development (i.e. the linguistic coverage and performance at a given point in time), but also to assess chances of carrying further the work done so far. This latter point requires very detailed knowledge of the internal structure of a system, e.g. of the system architecture, components, internal representations, etc.

A special case of development-related evaluation is the accompanying evaluation work running in parallel with system development. This can take place within a company or an institution developing a system or in collaboration with field validators, i.e. prototypical users or beta-testers of a given system who report regularly to the developer.[2] The interest here is in assessing the evolutionary development over a certain period of update cycles: the evaluation has to keep track of changes in the behaviour of the system, each time new elements (e.g. grammatical or lexical information, but also processing-related components) have been introduced. The practical work we have done on SYSTRAN falls into this area. We will give further details about the scenario in which we worked in section 3.

Finally, supervisors and sponsors of MT development, as well as researchers in machine translation and NLP in general, may have a more indirect, problem-oriented or efficiency-oriented view on the MT systems under evaluation. The system's capacity to cope with certain “hard problems” of translation may be more relevant, for example, to a researcher.[3]

Axes of evaluation

Different evaluation clients are interested in different aspects of a machine translation system and of its linguistic contents. From a more general point

[3] In German, for this group the term “Mitentwickler” has been coined.

[3] For a more detailed overview of types of evaluation clients and their respective needs, see the contributions to the panel at MT Summit III and King’s introduction to this panel.

of view, we distinguish four main axes along which MT systems may be assessed. Linguistic aspects are only one of these:

- *managerial aspects* include the organization of the machine translation facilities, general aspects of the use of a system in office communication and office organization, etc.;
- *ergonomic aspects* concern the technical side of user interfaces and of interaction with the system, the specifications of tool input and output, the amount of repetitive manual work necessary in the practical use of the system, the integration of the machine translation system into the office automation chain, practical aspects of the interaction and integration with other tools (e.g. steps to be performed in the work chain, file transfer, exchange of resources, etc.);
- *computational aspects* include the hardware/software platform, compliance with standards, usability in advanced computing environments, exchange of resources, data, etc. This area also covers "internal" aspects, such as the overall system architecture, types of internal representations, resulting extensibility behaviour, etc.
- *linguistic aspects,* finally, include the coverage of mono- and bilingual components, linguistic knowledge sources, linguistic processing, the extensibility of the linguistic components and possibilities of "system tuning", i.e. adaptability to particular linguistic needs.

Aspects of evaluation work

From the short overview given above, we can extract a number of dichotomies relevant for classifying evaluation activities. The following two are taken into consideration in the remainder of this paper:

- Snapshot evaluation vs. cyclic evaluation:

  — the purpose of snapshot evaluation is to get a picture of the behaviour of one or more systems at a given point in time; the evaluation action may concentrate on one system or compare several systems; it may concern one or several of the above axes of evaluation;

  - the purpose of cyclic evaluation is to follow in detail the changes of a system over a given period of its development, m order to check the system's performance over the update cycle against a (possibly constant) set of phenomena serving as an expectation horizon.

- General vs. application-specific evaluation:
  - general evaluation aims at verifying the overall performance of a system, taking as a starting point a subset of general language; mostly this subset is not determined by any specific requirements, but maybe by a general notion of "core phenomena", based on some set of criteria; the aim of this type of work is to test the system's performance with respect to a subset of phenomena determined on criteria which may be set up for the purpose of the evaluation, but which are not necessarily derived from the application profile;
  - application-specific evaluation is less interested in the performance of the system with respect to a wide variety of input texts, than in the question of knowing how well a specific system performs for the purpose of translating a given type of input material. This work will mostly be corpus-based, starting from a collection of the (different types of) texts most relevant to the client.

According to the degree of collaboration with the producer, and to the possibilities for the evaluator to get insight into the inner workings of a system, the two modes of "black box" and "glass box" evaluation have been distinguished in the literature. The usual snapshot situation will normally not allow any access to the internal representation used in a given system and thus be "black box". The most evident situation of a "glass box" evaluation is that of a developer who does accompanying in-house evaluation work, along with system development.

The distinctions we have discussed in the present paragraph (snapshot vs. cyclic, general vs. application-specific, black box vs. glass box) are cross-classifyable amongst each other and may in part introduce subclasses (e.g. a potential customer's snapshot evaluation of a given system may be general or application-specific).

3 Practical evaluation work with SYSTRAN French → German

We now describe the particular scenario in which we have been working when doing some partial evaluation work on SYSTRAN's French → German translation components. We first describe the situation in which our work was carried out, then give an overview of the criteria on which the test suites which we have used are based, and finally discuss a concrete example and the overall results of the work.

3.1 The SYSTRAN French → German scenario

Our scenario is a particular instance of "Mitentwicklung", i.e. a special case of accompaniment of a part of the system development cycle.

Since we are not system developers, we have in a way two "clients" or at least two discussion partners: the professional users and the developers. The professional users are the central services of the German Federal Railways Corporation (Deutsche Bundesbahn, Hauptverwaltung) in Frankfurt. They are conducting a study on the use of raw translation by specialized services of the German railways and of other railway companies; they are thereby in a position to collect end users' feedback on raw translation output of SYSTRAN. The developers are Telindus S.A., a software house working on SYSTRAN on behalf of the Commission of the European Community, DG XIII, at Luxembourg.

The topic of our evaluation work is limited to the linguistic performance of the SYSTRAN system's French → German translation. The evaluation is cyclic: we have followed, over two years, a number of development and update cycles of the system. The procedure consists of several steps: a number of linguistic phenomena are tested; the results are interpreted in cooperation with the developers and the users; if there are problems, the developers modify and update components of the system and then the impact of the modification is tested again by a second translation of the test material containing the relevant phenomena.

The evaluation has so far been mostly general: the users have to treat texts of highly diverging types, ranging from letters over technical and administrative memos to specialized articles and even newspaper articles. Consequently, since a common sublanguage could not be determined, we decided to first work on general language phenomena from grammar and lexicon which are of high frequency.[4]

The evaluation is not completely of the glass box type, although we have access to an informal description of the contents of the lexicons used by the system, in terms of an account of the types of linguistic information contained in these lexicons. No interpretation of internal representations produced by the system is carried out however, in this particular evaluation exercise. This

[4] See below, section 3.2.2 for details. We have described in more detail the different types of activities carried out in preparation for the work discussed here, in [HEID 1990] and in [HILDENBRAND/HEID 1991]. In the meantime, experience has shown that the impact of the improvements introduced by updates and modifications of the system would be greater for a more delimited type of input material. The end users have also expressed their interest in receiving above all raw translations of articles from specialized journals, under the condition, however, that the terminological quality of the translation results is acceptable. The goal for our further work with Deutsche Bundesbahn and SYSTRAN is consequently to profit from users' inhouse terminological resources in order to use the system as a means of quick access to domain-specific information.

is left to the system developers.

3.2 Test suites for French → German verb sub categorization

One of the fragment areas tested in the evaluation exercise described here is the use of lexical information about syntactic properties of verbs in the translation process. Test suites for this type of lexical and syntactic phenomena have been constructed and translated with the system.

3.2.1 Motivation

The objective of the partial evaluation work carried out so far was to come up with concrete proposals for improvements of SYSTRAN's treatment of basic high frequency grammatical constructions involving sub categorized complements of verbs. A need was felt for tests of this domain of the syntax and the lexicon, because numerous problems had been encountered in previous post-editing work; as a matter of fact, we have only started to work with test suites after a phase of almost two years of extensive post-editing and attempts at error classification.

The following are some of the problems encountered in postediting:

- Subcategorized prepositional complements were not recognized and/or translated by subcategorized complements. In some of the cases of complements realized by prepositional phrases, this problem could easily be identified because of the presence of default translations of prepositions in SYSTRAN: prepositions such as French *à, pour, par, sur*, etc. have a standard translation into German, mostly derived from the use of the preposition in local or directional adjuncts. The preposition *à* in its use as an indicator of an indirect object must thus in general be explicitly marked in the dictionary entry of the verb which takes this complement. Sometimes, this marking does not appear in the entries of the French verbs with indirect object encountered in the texts translated by the users, or the relevant information is not used in the translation process.

- German sentence generation has to keep track of word order constraints for placing the reflexive pronoun *sich,* etc. in those cases where a French verb is translated by a reflexive verb in German. Such cases, as well as those where French verbs are pronominal and German equivalents are not *(se promener ↔ spazierengehen)* necessitate a particular treatment in SYSTRAN's grammars and lexicons.

- If a verb has more than one meaning depending on the syntactic construction it takes, it often also has several translations depending on these meanings and their respective syntactic constructions, A typical example is the French verb *déduire,* which is equivalent to “infer” or to “deduct”, depending on whether it takes a propositional complement or not: *déduire du succès que la méthode est bonne* vs. *déduire 1.000 francs de la facture.* This phenomenon (polysemy, in a sense) needs to be kept separate from variation of realizations of a given verb argument which do not imply a major change of the verb meaning (and of the translation). A typical example of this latter phenomenon is the realization of propositional arguments by noun phrase complements, *that*-clauses or controlled infinitivals, respectively, as in *rêver de Marie, rêver de ce que Marie soit ici*, *rêver de partir pour les îles.*

  These two clusters of phenomena, polysemy of verbs manifesting itself by a change of the syntactic properties of different meanings, and variation in the realization of the complements of one and the same verb meaning must be adequately treated in a machine translation dictionary. The postediting situation suggested that the necessary separation was not always made in SYSTRAN’s lexical and grammatical knowledge sources. Thus, a more general approach to the problem than just postediting and local corrections was needed.

  Pairs of non-reflexive/reflexive verb constructions constitute a problematic case of the same type. A paradigmatic example of the polysemic case is French *prêter (qc) à qn* (“lend”) translated correctly by *jemandem (etwas) leihen* in German. However, *se prêter à qc* (“be apt”) should not inherit from the non-pronominal reading. But the SYSTRAN system does not always keep the two lexemes separate; consequently, *se prêter* sometimes receives the same equivalent as non-reflexive *prêter*: *se prêter à* is then wrongly translated by **sich jemandem/einer Sache leihen* instead of *sich für jmdn/etwas eignen.* We will come back to the distribution of the correct and the wrong translations later in section 3.3 and will use this example to illustrate our use of test suites.

After a fairly extensive postediting operation and subsequent local improvements made by the system developers on the basis of our suggestions, it was decided to go for a more systematic approach to the testing and update of the lexical and grammatical knowledge sources of the system. The reason for this is that many of the lexical problems encountered in the postedited texts were quite similar in type and therefore called for a more principled treatment. It was decided to use test suites for the detection of possible problems and for checking of the impact of the lexicon and grammar updates carried out by the system developers.

3.2.2 Principles of the construction of test suites for the example case

The basic principles of the test suites used in the experiments described here, are those laid down by [KING/FALKEDAL 1990]. In addition, a number of (fairly simple) rules were observed in the construction of the test suites used for SYSTRAN's French → German system. They concern the choice and form of the evidence submitted to the system and the evaluator's expectations in the analysis of the translation results.

The starting point for the construction of the test suites was an inventory of around 850 verbs, taken from SYSTRAN's monolingual French dictionaries and grouped according to the construction classes used in the system. The idea there was to follow the classification used by the system developers; the intention was to verify to what extent items classified syntactically as pertaining to the *same* construction class display the *same* behaviour in the processing, i.e. when the monolingual dictionary, the grammar rules and the translation rules are used together. No particular interest was devoted to the whys and hows of the classification, since the purpose was not to reinterpret or to reclassify the dictionary material. The only assumption was that the items taken by lexicon classes might cover a relevant range of different construction types described by the monolingual knowledge sources of the system.

An additional subclassification according to the types of modifications necessary under translation could have been used (e.g. structurally isomorphic translation; translation involving certain types of modifications, such as "change" of prepositions, change of argument-complement relations (the famous *like* ↔ *plaire* case!), etc.). Such a subclassification was in part used afterwards, in the analysis of the results, but it turned out not to be relevant in all cases.

Two types of test suites were produced, one which was thought of as containing only a very restricted set of linguistic phenomena ("basic test suite"), for each lexical item tested; the other set of materials contained the same sentences as the basic test suite, but with a number of additional phenomena.

3.2.2.1 The *basic test suite*: "trivial sentences"

For each item of the verb list, a "trivial sentence" in the present tense, affirmative form was constructed. These sentences form the *basic test suite.* The sentences are "trivial" in so far as the vocabulary used is very limited. The test suite was checked and amended several times, in order to make sure that

- all lexemes were translatable: besides the verbs which were tested, a very restricted set of nouns was used: e.g. *le chef, le secrétaire, le ministre* for person-denoting nouns, *le livre, le texte, l'hypothèse* for concrete and abstract object nouns, *la baignoire, la ville* for objects potentially denoting places, and so forth; the reason for this "minimalistic approach" was to avoid interference between lexical phenomena concerning the lexical material not really tested and the verbs under review;

- no categorial homographs (e.g. noun vs. verbal form) were contained in the test sentences;[5]
- no negation was contained in the test sentences;
- in a first version, only definite articles or demonstratives were used (as far as possible, with very few exceptions).

This basic test suite was translated several times with SYSTRAN, and a first, system-internal expectation horizon was constructed for this basic test suite: the system did not furnish "correct" translations for all sentences, in the sense that it did not produce grammatically acceptable German output for each such basic sentence. The results were classified into "acceptable" vs. "wrong" and then, for the "wrong" results, according to the problems occurring in the translation of the tested verbs. The major problem types were as follows:

- no German equivalent for the tested verb available;
- lexeme equivalent for the verb available, but the syntactic construction rendered in an ad hoc way (word by word, apparently without use of any subcategorization information which would have been similar to that apparently available for French);
- problems of German generation: word order in reflexive constructions, word order in sentences with complement clause, etc.

For subsequent testing, the test suite was separated into the subset of items correctly translated and the subset which had led to errors of the type explained above. The subsets were then both used for the creation of "derived test suites".

[5] Except *that,* complementizer or demonstrative, used only as a complementizer in our sentences.

3.2.2.2 "Derived test suites": only few parameter changes at a time

Once a stable result was obtained for the basic test suite, a number of modular variants of this test suite was produced. Modularity here means that, for each derived test suite, as few contextual parameters were changed as possible. So, for example, just the verb tense (past, passé composé, future tense) was changed, or an adjunct was added or the whole sentence from the basic test suite was embedded as a complement clause or a relative,

No lexical changes of the minimal vocabulary used in the basic test suite were made. Each parameter of change (tense, adding of adjunct, embedding, ...) was individually applied to all sentences contained in the basic test suite, thereby leading to several derived test suites equal in size to the basic one.

This approach allows, at least in principle, for the testing of individual grammatical parameters on top of the results of the basic test suite. Of course the choice to have the basic test suite in the present tense indicative, affirmative, is quite arbitrary. This choice was more motivated by practical than by any theory-driven considerations.

The parameters of variation introduced at this stage of the experiment were the following:

- tense: future tense and "imparfait" (past) in the French source sentences; the German translation of future tense requires the presence of an auxiliary (cf. *wird kommen,* etc.) which takes an infinitive. This has an impact on word order in German; "l'imparfait", on the contrary, leads to translations which are structurally identical to those of the sentences contained in the basic test suite;

- embedding in a *that*-clause *(le chef dit que* + basic sentence): relevant in German for word order (leading to verb-final, e.g. ..., *daß der Sekretär den Minister begrüßt,* vs. verb-second in *der Sekretär begrüßt den Minister);*

- embedding in relative clauses (e.g. *Je connais le chef qui* and *Je connnais l'entreprise dont le chef ...*): also relevant for word order in German *(ich kenne den Chef, der den Sekretär lobt* instead of *der Chef lobt den Sekretär,* etc.).

The following assumptions underlying the construction of these derived test suite modules were made, with respect to translation:

- None of the contextual parameters added (each parameter being tested with each of the 850 basic sentences) should affect the equivalence choice for the individual verbs, around which the basic test suite was built.

- The parameters interact however in part with the word order constraints for German, such that the derived test suites could be used to identify points in the grammatical and lexical programs of the system, where German generation failed to operate correctly. The operational goal was to allow for a partial update or modification by the system developers of the treatment of word order and constituent order in the German generation programs.

3.2.2.3 Subsets of test suites

Besides modularization at the level of the parameters added to the basic test suite contents, another type of modularization was used: partial test suites for subsets of the basic test suite were generated, e.g. for French reflexive and pronominal verbs, for verbs leading to German equivalents which take an obligatory expletive element *(der Chef rechnet* DAMIT, *daß der Sekretär kommt; *der Chef rechnet, daß ...*), etc.

The motivation for the extraction of such subsets of test suites is basically a practical one: the subsets are quite useful for the developer, once problems have been identified and individual types of constructions need to be treated. Within the work on German word order, the expletive elements were of major importance to the developers in view of a check of the German generation results, across or independently from the French constructions.

In other words: not only criteria of the source language, but also characteristics of the target language may play a role for breaking down the test suite into modules. Of course types of contrastive actions (structural isomorphism, differences in complement realization, etc.) would be another possible classification criterion.

Once the translation results for each of the derived test suites had been tested, a combination of the different grammatical parameters became feasible. So, for example, the basic sentences were tested not only separately under future tense and under embedding in *que*-clauses, but also in *que*-clauses and being themselves in the *future tense (le chef dit que le secrétaire viendra,* etc.).

This combination of several parameters allows to check whether the interaction of different phenomena has an impact on the translation results. Using this device, it was possible to identify unwanted interaction at a number of levels. In all cases attention was paid to avoiding any translational impact of the parameter combination; it was ensured (empirically) that the equivalence of the verb was not influenced by the combination of the parameters; an example of the type of interaction to be avoided is the use of negation and the modal "can" with a verb like *help: I cannot help doing* has a meaning of its own, different from the usual sense of *help.*

3.3 A concrete example and its interpretation

The construction of the derived test suites was meant to lead to a set of sentences differing among themselves only in a few parameters which would be "translationally irrelevant".

For a given verb, such a set of examples looks as follows:[6]

1. *Cette méthode se prête à ce problème.*
2. *Cette méthode se prêtera à ce problème.*
3. *Je connais cette méthode qui se prête à ce problème,*
4. *Le chef dit que cette méthode se prête à ce problème.*
5. *Je connais le chef dont la méthode se prête à ce problème.*
6. *Je connais cette méthode qui se prêtera à ce problème.*

The examples illustrate the following parameters:[7]

- (1): basic sentence: present tense
- (2): basic sentence + future tense
- (3): embedding in relative clause introduced by *qui* + present tense
- (4): embedding in an object clause under *dire que* + present tense
- (5): embedding under a relative clause introduced by *dont* + present tense
- (6): embedding in relative clause introduced by *qui* + future tense

The translation results expected would all have contained the German equivalent *sich eignen* for *se prêter.* For example, (1) should lead to *diese Methode eignet sich für dieses Problem.* No variation of the translation of the verb is expected, but variation is exactly what is found in the results produced by SYSTRAN: in most cases, *se prêter à* is translated wrongly as **sich leihen (+DAT),* only in the embedding under a *dont*-relative, the translation produced is *sich eignen für.* This result is not a single exceptional fact: there are some verbs for which the expected translation is produced in some syntactic contexts but not in others.

[6] The limited vocabulary mostly leads to artificially sounding sentences. However, these are easier to cross-check through large amounts of output.

[7] The variation could of course be continued, especially by combination of different parameters, such as embedding and tense of the embedded verb (illustrated in (6)).

We first try to interpret the individual case before trying to generalize. The result **sich leihen (+DAT)* is apparently constructed by use of “default” translations for the verb *prêter* which are based on the non-reflexive use. A sentence like *prêter 100 francs au chef* would be correctly translated as *dem Chef 100 Francs leihen.*

If no other result were produced, but just *sich leihen +(DAT),* the suggested interpretation would be to assume that the separate meaning of *se prêter à* and its translation were not available in the lexicons of the system. The fact, however, that *sich eignen für* is produced in the translation of (5) suggests that the particular translation of *se prêter à* is in principle available in the system, but not accessible in the contexts of (l)-(4) and (6). This in turn suggests that some unwanted interaction between the dictionary and the grammar rules takes place.

Similar effects were found in a number of cases; as well individual verbs as whole subsets of syntactic classes show this behaviour: a “default” translation is (wrongly) used in sentences with certain parameters, and a correct non-default translation in sentences with other parameters, parameter change being always translationally irrelevant. Another typical example is the translation of prepositional phrases which function as prepositional complements.[8]

We now have to interpret these results from the point of view of their implications for the use of test suites in evaluation. Again, we move from specific problems to more general ones.

First, recall that the “basic sentence” illustrating *se prêter* was wrongly translated (cf. (1), above). After the separation of the basic test suite items into “correctly translated” ones and “wrongly translated” ones, this sentence was in the second set. If this set had not been used to construct derived test suites, the above phenomenon would not have been detected, at least on this example. The use of all items from the basic test suite to construct derived test suites thus counterbalances in some way the a priori decision of the evaluators to arbitrarily use present tense affirmative sentences as “basic” ones (see above, section 3.2.2,2).

On the other hand, the example illustrates the need for a fine grained evaluation grid, in terms of the context types tested: if only the basic test suite or part of the parameters contained in the derived test suites had been tested, a less complete picture of the possible interactions of lexical and grammatical devices in SYSTRAN would have been visible, and still the picture is very incomplete.

At this point, of course, a distinction between procedural and declarative

[8] It has to be noted that some examples of this type can be interpreted as being ambiguous between a complement-PP and an adjunct-PP: *il pense à Paris* could be translated as *he thinks of/in Paris.* The examples of *prêter/se prêter* are more clearcut, although the PP-ambiguity problem may be solved in an MT system by deciding to give priority to translations involving fully instantiated subcategorization “frames”.

systems has to be made. In principle, one would expect a declarative system to produce more consistent results than a procedural system like SYSTRAN does. This is true at least for phenomena which can be treated compositionally and which the designers of declaratively oriented systems may have treated by composition of individual solutions for other phenomena. So, if an embedding under a relative clause and a change of the tense of the main verb can both independently be treated correctly by a given system, it is likely that the combination of both can as well. In this respect, SYSTRAN is maybe a particularly problematic candidate for testing, but the question still remains whether, for tests of a simple lexical phenomenon like the description of subcategorization properties of verbs, it is sufficient to use a very limited number of contexts in a test.[9]

So, even if a complex procedural system like SYSTRAN is an extreme case, the fundamental fact remains that test suites risk to grow rapidly in size. Given the size of the task, it seems all the more useful to stick to the concept of modularity and locality of parameter changes in the construction of test suites. The fact that the test suite is based on a minimal vocabulary, outside the verb lexicon tested, along with the concept of local parameter change makes the interpretation of the test suite results easier, since the test setup is more restricted to the facts under evaluation, and less “noise” is introduced. It may, however, be the case that this approach is somewhat biased to a test setup where the interaction of lexical and grammatical facts is tested.

4 Conclusion

We have presented an experiment on the use of test suites for the evaluation of part of the fragment coverage of SYSTRAN’s French → German translation module. The phenomena tested concern verb subcategorization and the translation of verbs in different contexts. The practical results show that even translationally irrelevant context changes, such as changes in embedding or changes of the tense of the main verb may affect the accessibility of lexical information in the SYSTRAN system. This in turn sheds some light on the need for large, fine grained and modular test suites, especially for evaluation of the linguistic performance of procedural systems.

[9] We cannot treat here the problem of testing the translations of those constructions licensed by the fact that a given verb is a member of a given syntactic class. For example, if a verb requires an indirect object, the French *à* + *NP* is pronommalized with *lui* in most cases; transitive verbs by definition can passivize, etc. To be sure about the availability of the full range of construction possibilities, pronominalization, passivization, etc, would need to be checked as well, depending on the syntactic classes. This step has not yet been undertaken in a systematic way in our SYSTRAN-related experiments.

The main outcome of this experiment is not so much in the facts we learned about SYSTRAN, but more in the methodological considerations which these results inspired:

- the basic test suite and the modules derived from it prove to be useful tools to keep track of the application of variation parameters to a large number of items (850 verbs, in the experiment);

- modularization of the test suites is necessary in order to avoid problems of the management of this evaluation tool. Test suites, in order to be fine grained enough, tend to be very large; the size and variety of parameters tested can be better managed if a strict module-wise structure is followed;

- even if the test suites are very large and detailed, it is still possible that some unforeseeable interaction between components, rules and data used in a system (especially a procedural one) cannot be captured; test suites thus may allow for a good approximation, still leaving open a number of untested cases.

If the experiment we conducted is in some sense characteristic for an evaluation of a commercial, procedurally implemented MT system, we may conclude from our experience that the effort which goes into test suite preparation, modification and update, as well as into the (manual) interpretation of the results is higher than what could be expected from an evaluator who wants to carry out a snapshot evaluation, especially an application-specific one. Such evaluation would require the setting up of new test suites for each application or client. This may turn out to be too costly for most clients. However, when it comes to evaluating the general performance of systems with respect to facts which need to be catered for in any MT system, reuse of test suites is possible and thus the effort may be more usefully spent. We consider the treatment of subcategorization in the verb dictionaries and the grammars to be such a "basic" type of phenomenon for which every MT system must have a solution. So, an evaluator may invest more in setting up test suites for this domain, of the grammar/lexicon interaction, if the material produced can be used to evaluate several systems. And again, snapshot evaluation may require other tools as well, since a test suite for the grammar/lexicon-interaction in the domain of verb subcategorization would not be sufficient to determine all of the linguistic performance of a system; it covers just one section of the fragment an MT system would have to treat.

From these considerations we conclude that the most efficient use we can see for the type of test suite we have described is in cyclic evaluation of a system under development. There, the same types of phenomena need to be checked after each update; furthermore, the possibility of extracting, from a given test suite, subsets according to different criteria (inspired by source language, target language or the contrastive treatment expected) makes it even more likely that such test suites will be best used for cyclic evaluation operations. Little effort had to be spent to extract specific subsets from the basic test suite and its derivatives.

References

[BARNETT et al. 1991] Jim Barnett, Inderjeet Mani, Elaine Rich, Chinatsu Aone, Kevin Knight, Juan C, Martinez (Microelectronics and Computer Technology Corporation, USA): "Capturing Language-Specific Semantic Distinctions in Interlinguabased MT"; in: MT Summit III Proceedings, July 1-4, 1991, Washington, D.C.; S. 25-32

[BATES 1988] Madeleine Bates: "*DRAFT CORPUS* for Testing *Natural Language data base* Query Interfaces", paper, distributed at the Natural Language Evaluation Workshop (Wayne, Pa.) 1988

[BILLMEIER 1982] Reinhard Billmeier: "Zu den linguistischen Grundlagen von SYSTRAN", in: *Multilingua 1-2*, 1982, S. 83-96

[FALKEDAL, K. 1990] Kirsten Falkedal: "Evaluation Methods for Machine Translation Systems: An Historical Overview and a Critical Account", Technical Report, ISSCO, Geneva.

[FELDBUSCH/POGARELL/WEISS 1991] Elisabeth Feldbusch, Reiner Pogarell, Cornelia Weiss: *Neue Fragen der Linguistik; Akten des 25. Linguistischen Kolloquiums, Paderborn 1990; Band 2: Innovation und Anwendung;* (Tübingen: Niemeyer), 1991

[FLICKINGER et al. 1987] Dan Flickinger, M. Friedman, M. Gawron, J. Nerbonne, C. Pollard, G. Pullum, I. Sag and T. Wasow: "HP-NL Test Suite", paper, distributed at the 1987 Linguistic Institute (Stanford, Ca.) 1987

[GAMBÄCK, B., ALSHAWI, H., CARTER, D. and RAYNER, M., 1991] Gambäck, B., Alshawi, H., Carter, D. and Rayner, M.: "Measuring Compositionality in Transfer-Based Machine Translation System", Workshop for Evaluation of Natural Language Processing Systems, University of California, Berkeley, California. (Also in this volume)

[HABERMANN 1986] Friedrich W.A. Habermann: "Provision and Use of Raw Machine Translation", World SYSTRAN Conference, 11.-14. February 1986 (Luxembourg)

[HABERMANN 1987] Friedrich W.A. Habermann: "Erfahrungen mit maschineller Übersetzung im Kernforschungszentrum Karlsruhe", Vortrag bei der Jahrestagung 1987 der Internationalen Vereinigung Sprache und Wirtschaft e.V., ms. (Karlsruhe), 1987

[HEID 1988] Ulrich Heid: “Maschinelle Übersetzungssysteme als Gegenstand der Übersetzerausbildung?”, in: *Fremdsprachen Lehren und Lernen (FLuL) 17 (1988),* 168 - 180

[HEID 1990] Ulrich Heid: Evaluation und Verbesserung der Sprachrichtung Französisch-Deutsch des maschinellen Übersetzungssystems SYSTRAN. Bericht des IMS für den Zeitraum 1.7.89-30.4.90 Internal Report, Stuttgart, 1990

[HILDENBRAND/HEID 1991] Elke Hildenbrand, Ulrich Heid: “Ansätze zur Ermittlung der linguistischen Leistungsfähigkeit von maschinellen Übersetzungssystemen – Zur Entwicklung von französisch-deutschem Testmaterial für SYSTRAN”; in: [FELDBUSCH/POGARELL/WEISS 1991]

[JIN 1991] Wanying Jin (New Mexico State University, USA): “Translation Accuracy and Translation Efficiency”; in: MT Summit III Proceedings, July 1-4, 1991, Washington, D.C.; S. 85-92

[KING 1989] Margaret King: A practical Guide to the Evaluation of Machine Translation Systems, Technical Report, ISSCO, Geneva.

[KING 1990] Margaret King: A Workshop on Evaluation: Background Paper. In: *Proceedings from the Third International Conference on Theoretical and Methodological Issues in* MT, pp. 255-259. Linguistic Research Center, University of Texas at Austin.

[KING/FALKEDAL 1990] Margaret King, Kirsten Falkedal: “Using Test Suites in Evaluation of Machine Translation Systems”, in: *Proceedings of COLING-90 (Helsinki) August 1990*

[KNOWLES 1979] F. Knowles: “Error analysis of *SYSTRAN-output* - a suggested criterion for the ‘internal’ evaluation of translation quality and a possible corrective for system design”, in: *Translating and the Computer (North-Holland Publishing Company) 1979,* S. 109-133

[LEICK/SCHROEN 1978] J. M. Leick and D. Schroen: Quelques résultats statistiques d’une évaluation sommaire du système de traduction automatique Systran. CETIL, CCE. Information document.

[LEHRBERGER/BOURBEAU 1988] John Lehrberger, Laurent Bourbeau: *Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation* (John Benjamins Publishing Company) 1988

[NIRENBURG (Ed.) 1987] Sergei Nirenburg (Ed.): *Machine Translation — Theoretical and methodological issues,* (Cambridge: Cambridge University Press) 1987

[ROUDAUD 1991] Brigitte Roudaud: A Procedure for the Evaluation and Improvement of an MT System by the End-User, talk presented at the *Evaluators’ Forum,* Les Rasses, 1991, this volume.

[SHIWEN 1991] Yu Shiwen: Automatic Evaluation of Output Quality for Machine Translation Systems, talk presented at the *Evaluators' Forum,* Les Rasses, 1991, this volume.

[VAN SLYPE 1976/78] Georges van Slype: "Zweite Bewertung des automatischen Übersetzungssystems *SYSTRAN* der Kommission der Europäischen Gemeinschaften für das Sprachenpaar Englisch - Deutsch. Entwurf" (Luxemburg: Kommission der Europäischen Gemeinschaften, Bureau Marcel van Dijk, Ingenieurs - Conseils en méthodes de direction) 1976/78

[VAN SLYPE 1979] Georges van Slype: "SYSTRAN, Evaluation of the 1978 version of the SYSTRAN English-French automatic system of the Commission of the European Communities", reprint from *The Incorporated Linguist,* Vol. 18, No. 3, Summer 1979

[WAY 1991] Andrew Way: A Practical Developer-Oriented Evaluation of Two MT Systems. University of Essex. Department of Language and Linguistics. Working Papers in Language Processing 26. June 1991