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

A First-Pass Approach for Evaluating Machine Translation Systems

Pamela W. Jordan
Mitre Corporation
C^{3}I Artificial Intelligence Center

1 Introduction

The MITRE Corp.[1] is in the midst of surveying and evaluating machine translation systems across the U.S. and, to a lesser extent, in Europe and Japan. The intent of the study is to recommend software purchases and R&D support that would meet the near-term, mid-term and long-term requirements of the users. Initially we identified over 20 machine translation efforts in the U.S. alone that we should investigate. Since it is too costly to do in-depth evaluations of so many MT efforts, we decided to gather just enough information to narrow down the possibilities. Once the best-fits according to our user's requirements are identified, deeper evaluations can be done on this smaller set. Planning and conducting the in-depth evaluations will take place at a later time. What is described here is our filter approach for narrowing down the possibilities and our assessment of its success to date.

The evaluation plan we formulated was to first predict what the most stringent requirements would be since, in our case, the requirements analysis time-frame overlapped with that of the survey. These requirements would determine what information was needed for the evaluation. Once the requirements analysis was complete, then the MT systems and research projects would be evaluated according to the projected near-term, mid-term and long-term needs of the users. The evaluation data collected would be prioritized and weighted based on these requirements so that all the relevant data could be combined to determine how well each system met the user's needs.

The requirements upon which we based our evaluation criteria were whether the MT systems and research projects provided the necessary functionality, whether the vendor's or research group's parent organization was stable enough financially so that we could reasonably expect them to continue their

[1] MITRE is an independent, not-for-profit organization that provides technical assistance, systems engineering, and acquisition support to U.S. government agencies

work and support the user, whether the system would be a good fit for the user's current and future concept of operations, whether it could be upgraded and maintained at reasonable costs, and whether it performed well enough to increase translation throughput. Figure 1 shows the mapping of these five broad requirements categories to the evaluation criteria we selected.

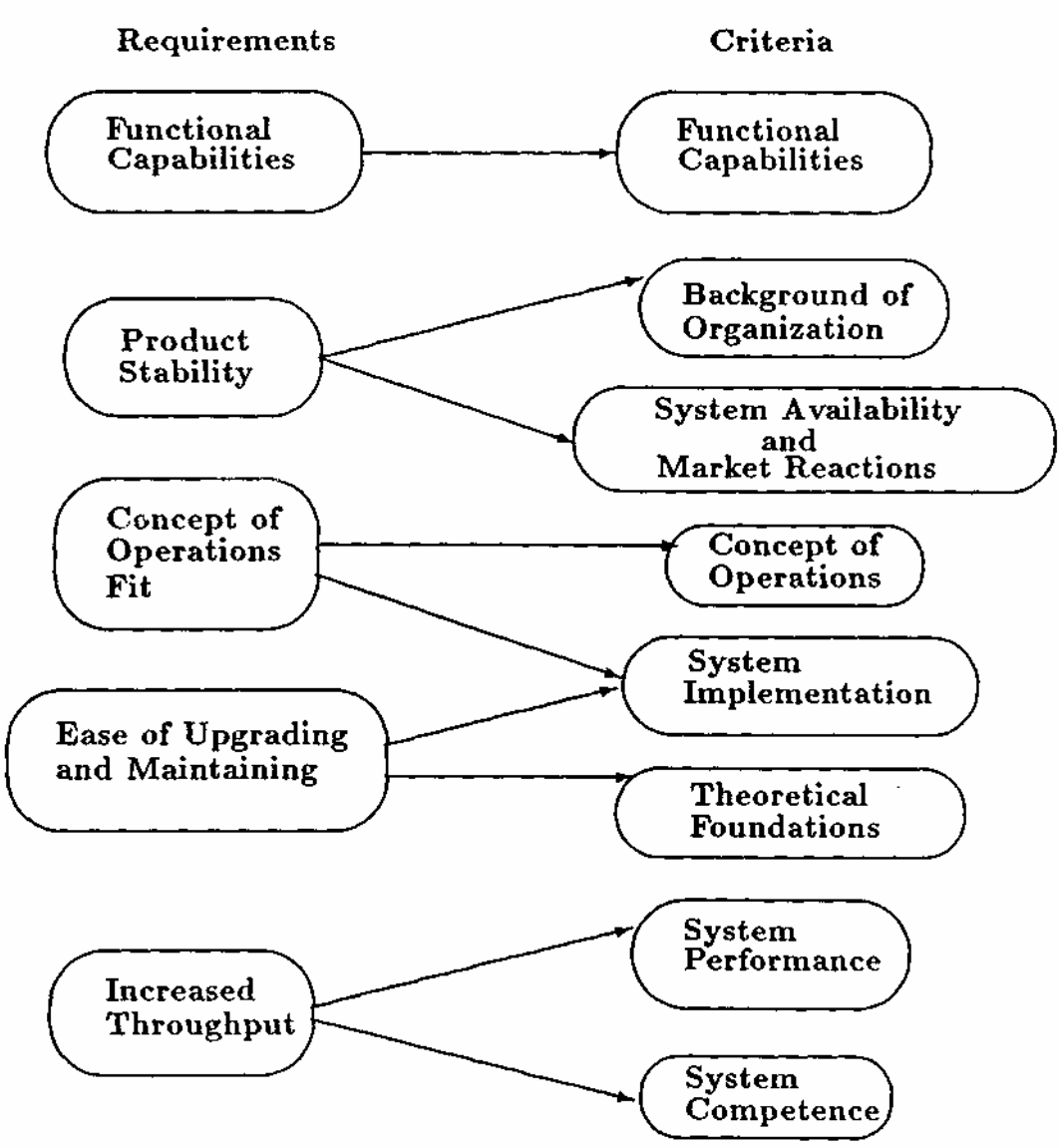

Figure 1: Mapping User Requirements to Evaluation Criteria

In describing the evaluation plan, I first discuss some of the current techniques for evaluating general software systems and natural language analysis systems, both of which are applicable to the MT evaluation problem, and I discuss current techniques and ideas for evaluating MT systems. Included in this discussion of evaluation techniques are the shortcomings of these approaches and an explanation of why it is so difficult to evaluate MT software. Next I describe our approach to the evaluation. This includes a discussion of the information to be collected for each of the evaluation criteria, as well as how this information is expected to contribute to the evaluation of the software for the near-term, mid-term and long-term requirements.

2 Evaluation Techniques

There has been a recent increase of interest in evaluating natural language systems, whether they are for translation, database stuffing or any number of other possible "NL-related" applications. This surge of interest is due to the growing popular opinion that natural language research has progressed enough to start transitioning some of it into applications software. Although the interest in evaluating MT software dates back to at least the ALPAC report (ALP66), the recent interest in evaluating other NL applications has sparked new discussion, techniques and insights for MT evaluation as well.

The approach one takes when evaluating software systems (in general) is two-fold: (1) evaluation of the accuracy of the input/output pairs; and (2) evaluation of the architecture of the system and the data flow between the system components. The former (external) view of software evaluation is equivalent to what NL researchers call "black-box" evaluation, and the latter (internal) view is referred to as "glass-box" evaluation (PF90). Black-box evaluation covers engineering issues such as reliability, productivity, learnability and likability (user friendliness). Glass-box evaluation also considers reliability (at the component-level) as well as maintainability, improvability, extendibility, compatibility and portability. (The engineering issues listed here are discussed in (GF88)).

NL applications have some unique problems that must be accounted for when doing black-box and glass-box evaluations. The biggest problem with evaluating NL applications is eliminating the subjectivity that, to date, has proven unavoidable due to the nature of natural language itself. Standard software engineering techniques for software evaluation apply to these NL applications, but these standard techniques must be enhanced to deal with the multiple "correct" answers that frequently occur with natural language. It is not clear what constitutes a correct answer especially when dealing with translations. Because of this, judging the correctness of the output for MT requires a degree of subjectivity. The output features to judge according to

Carbonell, et al. (CCG81), are semantic invariance (preservation of meaning), pragmatic invariance (preservation of purpose), structural invariance (preservation of syntax), lexical invariance (preservation of word senses) and spatial invariance (preservation of text format). Preserving the meaning of the source text is one of the predominant evaluation criteria used today in the MT research community.

Black-box evaluation, in the case of MT, tends to focus on evaluating the translation-quality of the output. Essentially it is an attempt to measure the acceptability of the translation to users. To produce the most objective measure possible, a standard test suite of input/output pairs should be established for judging whether the system is performing "correctly" or not and whether it will be cost effective. In light of the above discussion, this is a very costly undertaking and has yet to be satisfactorily accomplished in any evaluation of an MT system.

Another difficulty in developing black-box test suites is caused by the number of dimensions along which MT developers must limit their systems. These systems can be thought of as shells that are customized to apply to a particular domain, language pair, and type of text. To demonstrate these systems, the developers and researchers typically customize them to meet the needs of their largest projected market or to test the validity of a research hypothesis. The test suites must also be limited along the same dimensions, but then there is no common range among the systems. Because of this lack of commonality, some systems will need to be customized for the chosen ranges of the test suite no matter what ranges are selected. There is not much incentive to customize for evaluation purposes since a customer is typically expected to pay for customizations.

The glass-box approach attempts to evaluate the system's internal processing strategies to measure how well the system does something. According to the ideas for evaluating NLP systems (PF90), this type of evaluation should include a determination of the system's linguistic coverage, and an examination of the linguistic theories used to handle the linguistic phenomena. Determining the linguistic coverage means testing what linguistic phenomena are handled and to what degree. The examination of the linguistic theories used includes how closely these theories were followed in the implementation and noting what modifications had to be made to the theories. In addition, one should look at the performance of the system's various modules. The evaluation of each of these modules should be treated as individual black-box evaluations.

Again, test suites are often proposed as a way to determine a system's linguistic coverage. The difficulty here is the interaction between different linguistic phenomena (KF90). When creating the test suites, one must attempt to eliminate the interactions and test the smallest possible number of phenomena at one time. However, minimizing the interactions is difficult

and the test inputs grow quickly. To bound the problem, the test-suite developers must know what linguistic phenomena are of greatest importance to the users and be well-versed in linguistics and the languages of interest (KF90).

Test suites have also been proposed by (KF90) as a way to test the improvability of an MT system. Improvability tests assume that either the evaluator is working closely with the developer or that the evaluator is able to modify the system himself. The caveats mentioned earlier on bounding the problem, apply here as well.

3 Approach to Evaluation

Although we support the basic idea of black-box and glass-box evaluation that is being pursued for NLP systems, this survey was of such a short time frame that test suites could not be built nor could customized tests be performed. Our approach to evaluation (given the time restrictions) was to interview developers, researchers and current users of MT, attend MT demonstrations, survey the literature for additional details about the software, collect (for further evaluation) sample inputs and outputs for each language the software handles, and (whenever possible) arrange for evaluation copies of the software to be given to potential users.

A detailed questionnaire was developed to guide information collection during interviews, demonstrations and while reading the literature. Two questionnaires were developed; one directed at users and one directed at researchers and vendors. Both are found in Appendix 4.

In the next five subsections, I discuss the information we collected as it relates to the five requirements categories and our rationale for collecting this information. In addition, I will discuss how this information is expected to contribute to determining how well a particular system meets the requirements for each of the three time-frames.

3.1 Functional Capabilities

As mentioned earlier, to realistically tackle the language translation problem, developers of MT systems inherently impose limits along several dimensions. These dimensions include the domain, number of languages translated and types of text handled. This information is important in judging whether a system will cost effectively handle a particular user-group's near-term requirements. In addition, for near-term considerations, it is useful to know the sort of user expected in terms of skill level and the types of tasks for which the system was designed. This information allows us to judge how

well a system will fit into the user-group's operational needs. All of this information was gathered directly from the system developers.

3.2 Product Stability

An important factor in deciding whether to invest in research or in a product, is projecting whether the supporting organization is going to be available in the long-term to support their product or complete their research. One consideration in projecting the longevity of a group is the market reaction to any products they have produced. Another important factor in predicting the success of the group is whether they are qualified to do the work.

Background of Organization To get an idea of the group's MT credentials we wanted to know the group's technological heritage, the amount of research they planned to do and whether they planned to transition any of their research work. We were also concerned about the stability of the MT work and the stability of the infrastructure supporting this work. We needed to be reasonably satisfied that the group would be able to provide technical support in the long-term and that they could follow through in transitioning any research efforts. And finally, we wanted to know their goals and schedules for machine translation research and development.

System Availability & Market Reactions Referring back to cost effectiveness, another factor is the actual cost of the system and how much it typically costs to customize the system for languages, domains and text styles. We collected this information by asking the system developers to supply us with their cost data. We also wanted to know whether the system had been popular with those needing translation services. To this end we asked how many systems were in use and for information about their customers so that we could contact them.

3.3 Concept of Operations Fit

How well a system will fit in with the user's concept of operations depends on the concept of operations anticipated by the developers and on the operational requirements of the MT system.

Concept of Operations As mentioned in the functional capabilities section, we needed to know the anticipated external flow of the translation process. To judge how well a system would fit into the user-group's operational needs, we asked the developers what tools they provided for getting the text into the system, and for manipulating the source and target languages. We

asked to see demonstrations of these tools so that we could judge whether they appeared easy to use or if some specialized system knowledge was required to successfully use them. Also, we needed to know what was required of the user whenever the system encountered text it could not translate. The concept of operations information was gathered by questioning the developers and from seeing demonstrations of the system.

System Implementation To match against user requirements for operational environments, we had to know what the system's hardware and software requirements were. Preferably the hardware and software would be something the user already had available. Otherwise, the missing hardware or software should not be prohibitively expensive.

3.4 Ease of Upgrading and Maintaining

To ascertain the ease with which the software could be maintained and upgraded, we examined how it was implemented and the theoretical foundations upon which it was based. Because we could not build test suites that measure improvability (as described in section 4.1) or conduct formal tests at that time, this part of the evaluation is entirely subjective. As mentioned earlier, King and Falkedal reported an approach in COLING 90 (KF90) for an evaluation based on test suites that will be considered in the future when more time and resources can be allocated.

System Implementation We needed to judge how well the system was implemented to assure that maintenance and system enhancements would not cause problems later in the system's life-cycle. A well-implemented MT system should be designed as a shell, so that it can be readily customized for new domains, language pairs and text styles. In addition there should be tools to aid in customizing and modifying these areas. These tools were noted along with the targeted users for these tools. Demonstrations of these tools were requested as part of the criteria under concept of operations. Knowing what tools are available and who could best use them gave us an idea of whether the developer's services or other specialized services would be required to customize and later adjust the system. In addition to customization tools, the system should allow for core language knowledge to be separated from the domain dependent knowledge. For example, the core words of a language should not have to be re-entered whenever the system is customized for a new domain. The developers were asked if they had allowed for such modularity. Another factor in determining the quality of a system's implementation was whether the software could be easily maintained and integrated with other software. To determine how easily the software could be maintained,

we asked whether the software was modular and whether good documentation existed that accurately described each of the I/O modules. To determine if the system could possibly be integrated with other software, we asked if a programmer's interface had been developed and documented for the system.

Theoretical Foundations Knowing something about the underlying theoretics of a system gives us some indication of the facility with which it could be extended to handle new languages, domains and text styles. Past experience in software engineering tells us that ad-hoc systems not based on some coherent theoretical foundation are difficult to extend and maintain. Additionally, one of the objectives of the project is to identify research that looks promising for meeting long-term needs. Again, knowing the theoretical foundations of the research is a major factor in predicting whether the work is promising.

3.5 Increased Throughput

Judging whether an MT system performs well enough to increase throughput is a complex problem that depends not only on the speed with which the system produces a translation but also on the quality of the translation. To rate the quality of the output we must consider the types of text and domains the user needs translated, and the translation quality that will be acceptable to him. We knew that many different text styles, languages and domains were of interest to the users. So until their needs could be prioritized we considered what language phenomena out of all possible phenomena the MT systems and research addressed. Another factor in judging the system performance was whether the system would produce an output of high enough quality that the user would be inclined to post-edit it instead of doing his own translation directly from the source text and ignoring the MT output.

Performance We needed to determine whether the system operated efficiently enough and that the output quality was high enough to provide cost effective translations. To make this determination, we asked the developers for any performance measures they may have collected. These numbers did not allow us to make truly objective comparisons since they were made under different test environments. Because of this we asked the developers whether the performance figures were based on real text and whether the measures were based on use by system developers or translators who were not associated with the development effort. Knowing this information allowed us to compensate for the performance measures among different systems.

The performance figures we asked for were: the time spent pre-editing and post-editing, the speed with which the raw machine translation is pro-

duced and whether any customers have measured a change in productivity. Associated with the speed for producing raw translations, we asked how the handling of unexpected text affected this performance. In addition to these kinds of performance measures, we asked for any measures that have been made for how long it takes to customize or enhance the grammar, lexicon and domain.

Competence To get an idea of a system's linguistic coverage as part of judging its competence, we used a fairly comprehensive checklist of linguistic and textual phenomena which can be found in Appendix 4. The idea was to ask the developers which of these phenomena their system handles. We did not expect anyone to have achieved handling of everything on the list. Also we collected data on linguistic phenomena that are currently of low priority to the users but we expect this data to be useful in predicting whether the quality of a system's approach can be improved. At this point we want to be able to gauge how wide a system's linguistic coverage is. Presumably a wider coverage means that the system could produce a higher-quality output.

A major factor in judging a system's competence is whether the output is acceptable to the users. We have three types of MT users: those who need to scan material to estimate its relevance, those who want to know the content of the material and those who want publication-quality output. To determine whether a system's MT output will be acceptable to at least one of these three user groups, we will evaluate the English output of operational systems (see section 2 for discussion on using test suites for acceptability measures). The source texts used for this purpose include those that the developer typically provides as samples and an independently selected text that corresponds to the system's language pairs, domains and text styles. Whenever possible, the machine translation of the user-provided source text was performed in our presence. By being present during the machine translation, we know what modifications had to be made in order to get the output that we will be evaluating for acceptability. The modifications that were made (e.g. pre-editing, post-editing, lexical changes and additions) will give us more insight into the linguistic coverage of the system.

Two separate types of fidelity tests are needed to evaluate the acceptability of the output to all three types of users. The fidelity tests determine whether the meaning of the text has been retained in the translation (semantic invariance). Users who want to know just the subject area and users who need just the content of the text, are primarily concerned with semantic invariance; anything beyond this is a secondary consideration. On the other hand, users who want publication-quality output are concerned about stylistic and grammatical well-formedness as well as semantic invariance. For the latter two user groups, if incorrectly ordered words (along with other typical

errors) do not inhibit the user's understanding, then that user will be satisfied with the quality of the output. For example, most people would agree that the following sentence is still comprehensible despite the word ordering error.

*Incorrectly words ordered do not inhibit the reader's understanding.

The first fidelity test will be to examine the raw MT output and state the subject matter of the text without referring to the original source material. This test predicts whether the MT output will be acceptable to those who are scanning for texts in particular subject areas.

In the second fidelity test, we will compare the raw MT output to the original source material and rate how well the meaning of the original text was preserved. When rating the semantic invariance of the MT output, we follow a scale to keep the ratings consistent across languages and domains. We will be using Nagao's seven point scale for judging accuracy or fidelity ($N^{+}85$) (also like Van Slype's (vS82) measures of information transfer). Nagao's seven point scale follows:

1. Content of input sentence faithfully conveyed to output sentence. Translated sentence clear to native speaker and no rewriting needed.

2. Content of input sentence faithfully conveyed to output sentence and can be clearly understood by native speaker but some rewriting needed. Can be corrected by native speaker without referring to original.

3. Content of input sentence faithfully conveyed in output sentence but some changes needed in word order.

4. Content of input sentence generally conveyed faithfully in output but problems with things like relationships between phrases and expressions and with tense, voice, plurals and positions of adverbs.

5. Content not adequately conveyed. Some expressions are missing and there are problems with relationships between clauses, between phrases and clauses or between sentence elements.

6. Content not conveyed. Clauses and phrases missing.

7. Content not conveyed at all. Output not proper sentence; subjects and predicates missing. In noun phrases, main noun is missing or clause or phrase acting as a verb and modifying a noun missing.

This scale is for individual sentences, so the ratings for each sentence will be combined by taking the average. If the rating of the article's fidelity is in the range 1-4 then it is suitable for someone who wants to know just the content of the text or who wants to post-edit it. In contrast, a rating in the range of 1-5 would be suitable for scanning purposes. An option we considered was to use multi-lingual domain experts (instead of ourselves) to perform the fidelity tests described above.

An additional measure of acceptability was considered for those users who need publication-quality output. In this case, style and grammaticality are as important as semantic invariance. However, no one expects publication-quality translations to be automatically generated by an MT system without some form of human assistance. This assistance could come in the form of pre-editing, post-editing or interactive dialogs that occur during the translation process to help resolve difficulties as they are encountered. So the question in this case is whether the combined efforts of the post-editor and the MT system are more productive than the translator (and to be fair, any other MAT tools of his own choosing) without the MT system. To determine which is more productive, timings would need to be made for the two "configurations" in a well-controlled environment. For example, to make the measure objective, the translator and the post-editor should not be the same person and a statistically significant number of measures would have to be made. Given the time constraints of the project, this type of measure will not be done but it could be undertaken during a more detailed evaluation. For now, we are relying on any productivity measures users might be able to provide.

4 Status of the Evaluation

At this writing, we have surveyed most of the work we planned to cover and have collected the evaluation data described here. The requirements analysis is completed and we are in the process of weighting and prioritizing the evaluation data on the basis of what we have learned about the requirements.

At the beginning of the survey, we were initially concerned that the subjects of the MT interviews would not be willing to spend as much time as it took to go through our lengthy questionnaire. Fortunately, an overwhelming majority of the groups were extremely cooperative.

As we continue with the evaluation, we find that the data we collected have been adequate for doing a first-pass evaluation. The biggest difficulty we have encountered, so far, has been in finding out enough about the user's requirements to make an evaluation possible. Users cannot easily tell us what makes an MT output acceptable to them and this is one of the key elements in evaluating an MT system for our users.

Acknowledgements

I wish to acknowledge the "we's" I referred to throughout the paper. The other members of the survey and evaluation team are Dr. John W. Benoit, Adrienne J. Kleiboemer and George L. Marling, and our MT consultant is Dr. Bonnie J. Dorr from the University of Maryland.

APPENDIX: QUESTIONNAIRES

1 User Interviews

User Group Name:
Address:
System or Product used:

MT Requirements

Languages and directionality needed and which needs met:

Domains needed and which needs met:

Text styles needed and which needs met:

(S&T— legal— news— article titles & IR sentences— telegraphic)

Any other requirements met:

Any other requirements not met:

Interaction style (batch— interactive)

Target user (translator— non-translator)

Concept of Operations

Hardware platforms:

Memory requirements:

System extensibility, tools and targeted users for:
- Correcting grammar:
- Adding new languages:
- Adding/correcting lexicon: System modifiability
- Modularity of software:
- Modularity of lexicon:
- Modularity of grammar:
- I/O for modules well-documented?
- Hooks to integrate with other software tools (i.e. programmer intf.)?

What is the external flow of the translation process (user viewpoint):

How is source text input?

Are there tools to import machine-readable text (describe them)?

Describe Pre-editing capabilities:

Describe Post-editing capabilities :

Describe from the user viewpoint how unexpected text is handled (e.g. unable to translate):

Describe user-friendliness of tools (amount of training required):

Is there any consistency-checking for modifications?

Are there software hooks that allow the system to be integrated with other software?

System Performance

Performance figures based on real text?

Performance figures based on use by translators or system developers?

Average time spent pre-editing:

Speed with which produce raw translation (prefer words/hour):

Average time spent post-editing:

Increase in user productivity:

Time required to modify grammar:

Time required to modify lexicon:

Effect of unexpected text on system performance:

Benchmarks against translators of varying skill levels:

Time spent customizing and type of person responsible:

Time spent correcting and type of person responsible:

Time spent integrating with other software and type of person responsible:

System Competence

Give subjective estimate of translation quality:

Collect samples of source input and associated target output:

Size of lexicon:

Structure of lexicon, describe:
- Handling of synonymy:
- Handling of hyponomy:
- Handling of abbreviations and acronyms:
- Handling of numbers:

Size of grammar:

Checklist for Linguistic and Textual Phenomena Handled
- Lexical Ambiguity
- Lexical Selection
  - Lexical Divergences
  - Categorial Divergences
  - Conflational Divergences
- Syntactic Distinctions
  - Head Initial vs. Head Final
  - Null subject vs. overt subject
  - Free inversion vs. static positioning
  - Free word order vs. configurational
  - Limited movement vs. long distance movement
  - Thematic divergences
  - Structural divergences
- Quantifier Scoping
- Conjunction Scoping
- Negation Scoping
- Modifier attachment
  - Prepositional phrases
  - adjectives
  - adverbs
  - noun-noun compounds (idioms)
- Proper Nouns
- Auxiliary verbs
- Levels of embededness allowed

- Anaphora (inter and intra-sentential, bindings)
    - Pronominal
    - Prosentential
    - Proverbial
    - Proactional
    - Proadjectival
    - Temporal
    - Locative
    - Elliptic
- Metonomy
- Case Markings
- Tenses
    - Progressive
    - Present
    - Past
    - Future
- Aspect
    - Honorifics
    - Instantaneous vs. Over time
- Moods
    - Interrogative (intentionally left out tags and echoes)
        - wh-questions
        - alternative questions
    - Declarative
    - Imperative
    - Subjunctive
- Comparative constructions
- Relative Clauses (gap filling)
    - Restrictive Relative Clauses
    - Reduced Restrictive Relative Clauses
    - Non-restrictive Relative Clauses
- Text typography
    - footnotes & footnote references
    - headings and paragraph markings
    - word boundaries
    - line boundaries
    - punctuation
    - quotations
    - parenthetical insertions
    - tables
    - figures
    - formulae
    - units of measure (technical use kgms but non-technical don't)

2 Vendor/Researcher Interviews

Company Name:

Company address:

Marketing contact name and phone:

Technical contact name and phone:

System or Product Name:

General Capabilities

Languages currently supported and directionality:

Domains covered:

Text styles covered:

(S&T— legal— news— article titles & IR sentences— telegraphic)

Interaction style (batch— interactive)

Target user (translator— non-translator)

Target translation quality of output text:

System Implementation

Hardware platforms:

Memory requirements:

Software implementation languages:

Availability of source code:

System extensibility, tools and targeted users for:
Correcting grammar:
Adding new languages:
Adding/correcting lexicon:

System modifiability

Modularity of software:
Modularity of lexicon:
Modularity of grammar:
I/O for modules well-documented?
Hooks to integrate with other software tools (i.e. programmer intf.)?

Concept of Operations

Describe the anticipated external flow of the translation process (user viewpoint):

How is source text input?

Are there tools to import machine-readable text (describe them)?

Describe Pre-editing capabilities:

Describe Post-editing capabilities :

Describe from the user viewpoint how unexpected text is handled (e.g. unable to translate):

Describe user-friendliness of tools (amount of training required):

Is there any consistency-checking for modifications?

Are there software hooks that allow the system to be integrated with other software?

System Performance

Performance figures based on real text?

Performance figures based on use by translators or system developers?

Average time spent pre-editing:

Average speed with which produce raw translation (prefer words/hour):

Average time spent post-editing:

Average increase in user productivity:

Average time required to modify grammar:

Average time required to modify lexicon:

Effect of unexpected text on system performance:

Benchmarks against translators of varying skill levels:

System Availability and Market Reactions

Cost:

Describe customization services offered:

Average cost for customizing
- New language:
- New domain:
- New text style:

Number of systems in use:

Number of companies which have purchased:

Description of current customers:

Description of projected customers:

Is the Company stable?

System Competence

Give subjective estimate of translation quality:

Collect samples of source input and associated target output:

Size of lexicon:

Structure of lexicon, describe:
- Handling of synonymy:
- Handling of hyponomy:
- Handling of abbreviations and acronyms:
- Handling of numbers:

Size of grammar:

Checklist for Linguistic and Textual Phenomena Handled

- Lexical Ambiguity
- Lexical Selection
  - Lexical Divergences
  - Categorial Divergences
  - Conflational Divergences
- Syntactic Distinctions
  - Head Initial vs. Head Final
  - Null subject vs. overt subject
  - Free inversion vs. static positioning
  - Free word order vs. configurational
  - Limited movement vs. long distance movement
  - Thematic divergences
  - Structural divergences
- Quantifier Scoping
- Conjunction Scoping
- Negation Scoping
- Modifier attachment
  - Prepositional phrases
  - adjectives
  - adverbs
  - noun-noun compounds (idioms)
- Proper Nouns
- Auxiliary verbs
- Levels of embededness allowed
- Anaphora (inter and intra-sentential, bindings)
  - Pronominal
  - Prosentential
  - Proverbial
  - Proactional
  - Proadjectival
  - Temporal
  - Locative
  - Elliptic
- Metonomy
- Case Markings
- Tenses
  - Progressive
  - Present
  - Past
  - Future

Aspect
- Honorifics
- Instantaneous vs. Over time

Moods
- Interrogative (intentionally left out tags and echoes)
  - wh- questions
  - alternative questions
- Declarative
- Imperative
- Subjunctive

Comparative constructions

Relative Clauses (gap filling)
- Restrictive Relative Clauses
- Reduced Restrictive Relative Clauses
- Non-restrictive Relative Clauses

Text typography
- footnotes & footnote references
- headings and paragraph markings
- word boundaries
- line boundaries
- punctuation
- quotations
- parenthetical insertions
- tables
- figures
- formulae
- units of measure (technical use kgms but non-technical don't)

<u>Descriptions of Theoretical Foundations</u>

MT approach (direct—transfer—interlingual—example-based)

Morphology, describe:

Morphological approach used (e.g. KIMMO) including changes and assumptions:

Heuristics used:

Representational mechanisms:

Syntax, describe:

Syntactic Theory used (GB—LFG—CUG—GPSG— HPSG— TG—other)

including changes, assumptions and heuristics:

Grammar design (corpus-based—standard grammar):

Parser used including changes, assumptions and heuristics:
Generator used including changes, assumptions and heuristics:
Representational mechanisms:

Semantics, describe:
Semantic Theory used for lexicon, world knowledge and domain knowledge including changes, assumptions and heuristics:
Representational mechanisms:

Discourse, describe:
Discourse approach used including changes, assumptions and heuristics:
Representational mechanisms:

Describe Architecture of the MT components (how do the above components interact?):

Describe approach to handling unexpected text:

Describe approach for consistency-checking:

List relevant technical literature that describes the theoretical foundations:

Of the linguistic phenomena handled, how is it done?

Company Background and Information

What is the technological heritage of the group?

What is the corporate structure and how does the MT group fit in this structure (is the support for MT serious)?

What are the goals and schedules for:
Commercial MT Software?
MT Research?

References

ALPAC. Language and machines: Computers in translation and linguistics. Technical report, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC, 1966. A Report by the Automatic Language Processing Advisory Committee.

Winfield S. Bennett and Jonathan Slocum. The LRC machine translation system. *Computational Linguistics,* 11:116 - 118, April-September 1985.

Jaime G. Carbonell, Richard E. Cullingford, and Anatole V. Gershman. Steps toward knowledge-based machine translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-3(4), 1981.

Tom Gilb and Susannah Finzi. *Principles of software engineering management.* Addison-Wesley Pub. Co., Reading, Mass., 1988.

Pierre Isabelle and Laurent Bourbeau. TAUM - AVIATION: Its technical features and some experimental results. *Computational Linguistics,* 11:24 -26, January-March 1985.

M. King and K. Falkedal. Using test suites in evaluation of machine translation systems. In *COLING 90,* volume 2, pages 211 - 216, 1990.

M. King. A practical guide to the evaluation of machine translation systems. Technical report, ISSCO, 1989. Intermediate Report to Suissetra.

J. Lehrberger and Laurent Bourbeau. *Machine Translation: Linguistic Characteristics of Machine Translation Systems and General Methodology for Evaluation.* Benjamins, 1988.

Makato Nagao et al. The Japanese government project for machine translation. *Computational Linguistics,* 11, April-September 1985.

M. Palmer and T. Finin. Workshop on the evaluation of natural language processing systems. *Computational Linguistics,* 16(3), 1990.

Muriel Vasconcellos and Marjorie Leon. SPANAM and ENGSPAN: Machine translation at the Pan American Health Organization. *Computational Linguistics,* 11:135, April-September 1985.

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