[*Progress in Machine Translation*, ed. Sergei Nirenburg. Amsterdam: IOS Press, 1993.]

5 Evaluating machine translation systems in a real work environment

• DORIS ALBISSER

As a potential customer of MT-software, I would evaluate commercially available systems using the following approach, which reflects the strategies employed at UBS:

When evaluating a machine translation system for productive use within a company, the underlying principle is to evaluate it as an overall system and not only for the quality of the MT-output. Also, the evaluation criteria should be designed so that they provide a true basis for comparison.

In this respect, the evaluation criteria to be taken into account can be subdivided into four main categories:

- the linguistic capabilities of an MT-system
- the technical environment provided
- the organizational changes involved
- the corporate situation of the MT-supplier

As a preliminary remark, it should be pointed out that the evaluation criteria and, in particular, their weighting are company-specific and thus subjective to some extent. Furthermore, quality issues are not quantified, they are rated according to their degree of importance. As regards the procedure, our evaluations are carried out inhouse using company- specific texts in both economics/banking and information technology (e.g. user manuals). This approach has proven worthwhile since it allows testing of the MT-system in the actual environment and not in a demo-room outside the company. Below illustration of the four main categories briefly outlines the evaluation strategies employed.

As for the linguistic part, we have developed a method to assess the quality of the raw MT-output. First, the sentences of a given text are categorized according to their degree of complexity ranging from I to IV (see Table 1). Second, the mistakes found in the raw translation are scored according to the criteria listed in Table 2. Basically, the determining factors for scoring the mistakes are whether they can easily be corrected, whether they

seriously hamper understanding, and whether they violate basic grammatical structures. It should be noted that the linguistic evaluation is largely language-dependent and to some extent even specific to the text type. The model shown in Tables 1 and 2 was designed for translations from German into English. The systems we evaluated with this method are TSS from LOGOS and METAL. TOVNA MTS had to be evaluated differently since German/English was not yet included in the translation versions available at that time.

Table 1: Classification of Test Sentences in the Source Language

Sentence Category I:	0-4 Points	Simple Sentence
Sentence Category II:	5-9 Points	Sentence of medium complexity
Sentence Category III:	from 10 Points	Sentence of high complexity
Sentence Category IV:		Sentences with syntax errors or sentences with a disproportionate number of potential errors

Further, the technical environment offered by the MT-software supplier has to meet certain requirements so as to comply with the corporate information technology strategy (e.g. open systems architecture). This may include portability, interfaces to sophisticated word processors (WPs) and desktop publishing systems, access to terminology from the WP mode, import/export of terminology, options for information retrieval (e.g. for recurring texts, updates), single vs multi-user system, and — most important — user-friendliness. Another important factor is the system's capabilities for further enhancement. Since commercially available systems tend to lend themselves to specific text types, the question arises to what extent an MT-system could be customized to meet the user's needs to optimum effect. Finally, in view of future integration of MT-systems into a corporate environment two general questions might be worth a moment of reflection. First, what is the potential of an MT-system to be integrated into a translator workstation? Second, does the MT-supplier take into account that translation is only part of the entire document production process or does he offer the MT-system as an isolated component?

The third and very often neglected evaluation criterion refers to the organizational changes involved. An evaluator has to determine the required user profile. In this context some issues must be clarified: Are terminologists needed? Who administrates the system? Can presently employed translators be trained (and if so, what is the learning curve)? Another important factor comprises the cost/benefit analysis. Thus, what is the price of the system, what is the minimum translation volume to justify MT, and what is the throughput per day, including both the volume of MT-output and the time required for dictionary coding and pre-/post-editing. The latter can only be estimated. As a consequence, the increase in productivity can be assessed during the evaluation phase to a limited extent only.

Finally, the corporate situation of the MT-supplier plays an important role in terms of future development and cooperation. Issues raised in this respect include the size of the company (resources for development), the importance of MT-software within the overall product range, the market share, management, the financial situation, and — very important — customer support. Close cooperation with the MT-supplier is crucial during an evaluation for it allows the potential MT-user to establish a sound business relationship which in turn

Table 2: Error Types

Error Category I:	**1 Point**
Determiners Pronouns Prepositions Reflexive verbs in German → Passive voice in English	
Error Category II:	**2 Points**
Adjective as present participle Pronoun agreement Adverbial word order: manner - place - time Prepositional objects Adverbs	
Error Category III:	**3 Points**
Verbs (examples of possible mistakes): • transitive, intransitive, reflexive • Auxiliary verbs • Active - passive voice • Incorrect tenses • Imperative form • Verb + preposition Nouns: ▪ Noun + Preposition • Genitive Attribute • Nouns in singular / plural only Adjectives: • Comparative • Predicative Word Order (E): S-V-DO-IO Lost parts of sentences Unrecognized words	
Error Category IV:	**4 Points**
Choice of Words/Lexicon (dictionary errors despite prior coding)	
Error Category V:	**5 Points**
Nontranslated sentences	

facilitates detecting the customer's needs for customization. Again, it is indispensable for evaluators to specify and communicate their corporate requirements to MT-suppliers if future systems are to be enhanced and tailored to individual needs.

In conclusion, I would like to emphasize that all four parts (linguistic, technical, organizational, and supplier-related) are of equal importance. Thus, an MT-system is evaluated as an *overall system.* Finally, subjectivity cannot be avoided in an evaluation because each company has its own needs and priorities. What might be generalized to some extent is the evaluation criteria as such, but not their weighting.

Below is a brief outline of how the complexity of a sentence is assessed:
A sentence with subject, verb, object and adverb is considered a standard sentence (0 points). Each additional part of a sentence is given 1 point. Furthermore, the subordinate clauses and individual parts of speech are scored as follows:

Subordinate clauses:	Points
1) with finite verb:	
in a clause	3
preceding a clause	2
following a clause	1
2) with non-finite verb:	
in a clause	4
preceding a clause	3
following a clause	2
Specific clauses:	
non-defining relative clause	2
subordinate clauses introduced by "and/or"	3
Main clauses:	
each additional complete main clause	1
incomplete main clause (criterion: at least a verb and a noun phrase)	2

Parts of speech (this list is incomplete as it would be too extensive to specify each item):

Verbs:	
conditional tense, indirect speech, imperative	1
impersonal use of "sollte, dürfte, etc."	3
Pronouns:	
reflexive pronouns	1
Attributes:	
Genitive attribute	2
Attributes with participle constructions	5
Each additional part within attribute	2
Sentences without subject:	3

Again, the purpose of illustrating our error allocation is to show what aspects may be taken into account when assessing the MT-output. It is not intended to give a full account

of each detail. As for the scoring, error category 1 shows minor errors whereas category 3 shows the most serious errors. Dictionary errors are counted separately (category 4). The same applies to non-translated sentences (category 5).