
*Machine Translation* **16:** 175–218, 2001.



# MTranslatability

ARENDSE BERNTH and CLAUDIA GDANIEC
*IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA*
*E-mail: {arendse,cgdaniec}@us.ibm.com*

**Abstract.** Machine Translation of arbitrary input is difficult, but the output quality can be improved significantly if writers create documents with MT in mind. This article deals with "MTranslatability" – translatability of texts by MT systems. It identifies characteristics of text that decrease MTranslatability and suggests ways to improve them. It also illustrates the effect of writing for MTranslatability by showing before-and-after pictures of output from various commercially available MT systems, and gives an overview of tools that help identify and correct the problems.

**Key words:** controlled languages, pre-editing, pre-editing tools, semantic annotation, translatability, translation confidence

## 1. Introduction

Machine Translation (MT) output is rarely perfect in its raw state. For most uses, the output needs to be post-edited. Of course, output quality varies from system to system, from language pair to language pair, and from domain to domain. Every MT developer knows how difficult it is to produce the desired high output quality for arbitrary input. And quite often, when we see the input that MT systems are supposed to be able to handle, it strikes us how much a few simple measures would help. If only the writer had avoided an ambiguous construction or had been more careful with the mark-up! If only the writer had not misspelled *they're* as *there*! In fact, if only the writer had been aware of the limitations of current MT systems and had created the document with "MTranslatability" in mind. By "MTranslatability" we mean, obviously, translatability as it relates to MT.

This paper is an attempt at a systematic analysis of various problems that MT systems encounter in free-form documents. It describes ways to rectify the problems, and tools to help identify and correct them. Many people have written about good writing, including writing for MTranslatability; however, as far as we are aware, a comprehensive overview has hitherto been lacking.[1] We hope that this paper will be of value to both users and developers of MT. Increasing awareness of how to write for MT should contribute to better user acceptance of MT once the users learn a few tricks that will enhance the output quality. And the paper can be seen as a summary of some challenging problems that MT developers need to address.

A different approach to improving MT output quality is that of user guidance during translation, or "Interactive MT" (IMT). The interaction traditionally relates to the source text, and in most cases the interaction with the user takes place as a written dialog, with the MT system in control. The idea was first introduced for the MIND system (Kay, 1973). Understandably enough, the concept became very popular with MT developers, and a number of IMT systems were developed, e.g. for translating between English and Japanese (Whitelock et al., 1986; Tomita, 1986). Maruyama et al. (1990) describe a somewhat different approach for Japanese-to-English IMT where the interaction is graphical, and the user mouse-clicks to change the dependency structure provided by the syntactic analysis, if necessary.

Langlais et al. (2000, 2002) propose a completely different view of IMT, used for the TRANSTYPE system. The focus is shifted from analysis of the source text to the *form* of the target text. TRANSTYPE, which uses a statistical MT system, provides guesses at text completion for the human translator, and is perhaps better characterized as a Machine-Aided Human Translation System.

IMT as a way of improving the output of MT is most suitable for professional translators who know both source and target language. IMT can also be useful for casual email where the sender and receiver are not familiar with each other's language. However, as more and more MT systems become available on the Web for the casual user who wants to take advantage of the large amount of information posted there, IMT is not always a viable approach. Anybody who makes documents available on the Web must be prepared for the possibility that somebody applies MT to their text without much (or any) knowledge of the source language.

Even in cases where the MT user has good knowledge of the source language and has the freedom to edit the source text, as is typically the case when the user is a large company, it often makes more sense to write with MT in mind from the beginning. This scenario also often involves translating the same source text into *several* languages, and it is more cost-effective to eliminate as many problems as possible in one source text than during several different translation processes. Mason and Rinsche (1995) make the point that the more the source text can be designed and created with translation in mind, the less work it will require to translate the document.

This idea is the basis of the concept of Controlled Languages (CLs), yet another well-known way of improving MTranslatability. One of the earliest examples of combining strict control of the input text with MT is the Xerox Multinational Customized English (Elliston, 1979), used with Systran. However, a major difference between our recommendations and the CL approach is that MT-oriented CLs mostly are tightly tied in with a *specific* MT system, whereas our recommendations can be viewed as more general.

It is also possible to combine the approaches of CL and IMT to ensure high-quality MT output, as is the case for the TITUS system (Ducrot, 1989), designed by the Institut Textile de France.

In this paper we shall identify some rules, which, if followed, will improve MTranslatability. Our goal here is not to address the issue of how best to present the rules to the writers.

In Section 2 we describe ways of writing documents to enhance MTranslatability. We look at grammar, ambiguity, style, punctuation, spelling, and mark-up. We give examples of their effects on MT output by showing before-and-after pictures of output from various commercially available MT systems. The examples all show English as the source language, but the general principles carry over to other languages.

In Section 3 we give an overview of the various types of tools that can help the user in the process of improving MTranslatability. These tools include spell checkers, grammar and style checkers, CL checkers, and annotation editors. We evaluate the relevance of each type of tool to MTranslatability and give examples of commercially available systems or research versions. In Section 4 we describe two approaches to automatic measurement of MTranslatability. Finally, in Section 5 we summarize our findings about what the writer can do to improve MTranslatability and give a brief outline of where research in MT needs to go in order to reduce the need for writers to write specially for MT. Here we also address the issue of evaluating our recommendations. Appendix A provides a list of all the rules given in the paper. In Appendix B we give an example of application of the rules to German and the effect on German⇒English translation. Appendix C contains a similar example for English⇒French translation.

## 2. Ways to Improve MTranslatability

In this section we describe a variety of issues that interfere with MTranslatability and offer guidelines for improving documents intended for MT.[2] Section 2.1 describes some problems that arise from the interaction of syntax and semantics. In Section 2.2 we address problems related to ambiguity. Section 2.3 looks at style issues, and in Section 2.4 we consider punctuation. Section 2.5 deals with misspellings and user dictionaries. Finally in Section 2.6 we address problems arising from basic file characteristics, such as improper use of mark-up.

The current section contains many recommendations. We realize that it is most likely unrealistic to assume that *all* recommendations can be followed, but our goal is to make the list as comprehensive as possible. The better the adherence to the rules, the better the chance of reasonable MT output. Also, there may be more than one way to handle a given problem. For example, it may be better to use mark-up (annotation) than rewriting in certain cases, e.g. coordination, in order to preserve idiomatic English. But annotation is still in its infancy, so it seems prudent to mention the possibility of rewriting. Given these recommendations, the authors (or whoever is responsible for the quality of the document) will be able to make their own decisions about what is feasible and what is not feasible.

It is also worth mentioning that some rules might seem contradictory. For example, a fair number of rules demand that the writer be more explicit, and the recommendation is to add disambiguating words or otherwise be more verbose. In certain cases this may pose a conflict with the rule of avoiding too long sentences, and human judgment is definitely called for.

### 2.1. CHECK THE GRAMMAR

*Rule 1.* Avoid ungrammatical constructions.

Current commercial MT systems rely on syntax to a large extent; therefore, ungrammatical input often produces wrong output.

Ungrammatical input can produce undesired results in more than one way. The most obvious problem is that the segment will not parse. A partial or wrong parse will spill over into subsequent processing steps.

We use "grammaticality" to mean correct grammatical form *for the intended meaning*. There are many cases of problems where a segment makes syntactic sense, but does not have correct syntax for the intended meaning. The sentence in example (1) occurred in a mail-order catalog; the translation to German, also shown in (1), is clearly not good.[3]

(1) *Woven of combed cotton, you will love our sweater's soft feel.

*Wenn Sie von gekämmter Baumwolle gewoben werden, werden Sie*
IF YOU OF COMBED COTTON WOVEN ARE WILL YOU
*das weiche Gefühl unseres Pullovers lieben.*
THE SOFT FEEL OF-OUR SWEATER LOVE

'If you are woven of combed cotton, you will love the soft feel of our sweater.'

Humans will usually ignore the fact that what the sentence *really* says is that *you* are made of combed cotton. However, most MT systems will have to go by the grammatical rule that the subject of *woven* has to be the same as the subject of the main clause, *you*. This can give funny results for target languages such as German that do not have this subjectless nonfinite construction but need a finite construction with an explicit subject.

If the sentence is rewritten to conform to grammatical English, this problem does not arise. Example (1) can be rewritten as either (2a) or (2b), with consequent improvements, if not perfection, in translations.

(2) a. Woven of combed cotton, this sweater will delight you with its soft feel.

*Wenn dieser Pullover von gekämmter Baumwolle gewoben wird,*
IF THIS SWEATER OF COMBED COTTON WOVEN IS
*wird er Sie mit seinem weichen Gefühl erfreuen.*
WILL IT YOU WITH ITS SOFT FEEL DELIGHT

'If this sweater is woven of soft cotton, it will delight you with its soft feel.'

b. Our sweater is woven of combed cotton, and you'll love its soft feel.

*Unser Pullover wird von gekämmter Baumwolle gewoben und Sie*
OUR SWEATER IS OF COMBED COTTON WOVEN AND YOU
*werden sein weiches Gefühl lieben.*
WILL ITS SOFT FEEL LOVE

'Our sweater is woven of soft cotton, and you will love its soft feel.'

Similarly subjectless nonfinite clauses involving a present participle as in (3a) should be written as (3b).

(3) a. *After inserting the diskette, the system will read the file.

*Nachdem das System die Diskette einführt, wird es die Datei*
AFTER THE SYSTEM THE DISKETTE INSERTS WILL IT THE FILE
*lesen.*
READ

'After the system inserts the diskette, it will read the file.'

b. After you insert the diskette, the system will read the file.

*Nachdem Sie die Diskette einführen, wird das System die*
AFTER YOU THE DISKETTE INSERT WILL THE SYSTEM THE
*Datei lesen.*
FILE READ

'After you insert the diskette, the system will read the file.'

## 2.2. REDUCE AMBIGUITY

Ambiguity is detrimental to MTranslatability. If a word or construction is ambiguous, then the MT system must try to determine the intended meaning, and this is as likely to come out wrong as right. It is better that the writer makes the decision before the text is machine-translated. There are a number of ambiguous constructions that can easily be written in an unambiguous way if the writer takes a few rules into consideration. In some cases, the resulting English may not seem quite as idiomatic; whether this is acceptable is a matter of policy or taste.

Often, structural ambiguity is caused by a telegraphic style and can be removed by use of a fuller style. Kohl (1999) calls this "using syntactic cues".

In Section 2.2.1 we look at coordination. Section 2.2.2 addresses problems caused by *ing*-words. Section 2.2.3 describes problems relating to postnominal modifiers. In Section 2.2.4 we take a look at pronouns, and in Section 2.2.5 we briefly mention various other problems.

### 2.2.1. *Coordination*

Coordination is a particularly difficult construction for MT to handle because of potential ambiguity in the scope of modifiers.

*Rule 2.* Repeat final words of the left conjunct or initial words of the right conjunct, as necessary, to disambiguate the coordination.

In (4) we give an example of a multiply ambiguous construction that involves coordination, and we show its translation into German.

The interpretation chosen by the MT system that translated (4) is shown by brackets in the English gloss at the end of (4). If this interpretation was the intention, the English could be written unambiguously by distributing *with the* and *responses* as shown in (5).

(4) The application can use the window to establish a dialog with the user and format text responses.

*Die Anwendung kann das Window benutzen, um einen*
THE APPLICATION CAN THE WINDOW USE IN-ORDER-TO A
*Dialog mit den Benutzer- und Formattextantworten*
DIALOG WITH THE USER- AND FORMAT-TEXT-RESPONSES
*herzustellen.*
TO-ESTABLISH

'The application can use the window to establish a dialog with [the [user and format text] responses].'

(5) The application can use the window to establish a dialog with the user responses and with the format text responses.

*Die Anwendung kann das Window benutzen, um einen*
THE APPLICATION CAN THE WINDOW USE IN-ORDER-TO A
*Dialog mit den Benutzerantworten und mit den*
DIALOG WITH THE USER-RESPONSES AND WITH THE
*Formattextantworten herzustellen.*
FORMAT-TEXT-RESPONSES TO-ESTABLISH

A different interpretation of (4) can be written unambiguously as in (6). Here the left conjunct is *with the user* and the right conjunct is *with the format text responses*. The disambiguation is accomplished by repeating *with the* in the right conjunct.

(6) The application can use the window to establish a dialog with the user and with the format text responses.

*Die Anwendung kann das Window benutzen, um einen*
THE APPLICATION CAN THE WINDOW USE IN-ORDER-TO A
*Dialog mit dem Benutzer und mit den*
DIALOG WITH THE USER AND WITH THE
*Formattextantworten herzustellen.*
FORMAT-TEXT-RESPONSES TO-ESTABLISH

The conjunction in (4) could also coordinate verb phrases. Examples of this are shown in (7a) and (7b). The left conjunct could be the verb phrase *use the window to establish a dialog with the user* and the right conjunct could be the verb phrase *format text responses*. This can be written unambiguously by repeating *can* in the right conjunct as shown in (7a). Yet another possibility is that the left conjunct is the verb phrase *establish a dialog with the user* and the right conjunct is the verb phrase *format text responses*. In this case, it would be better to distribute *to* as shown in (7b).

(7) a. The application can use the window to establish a dialog with the user and can format text responses.

*Die Anwendung kann das Window benutzen, um*
THE APPLICATION CAN THE WINDOW USE IN-ORDER-TO
*einen Dialog mit dem Benutzer herzustellen und kann*
A DIALOG WITH THE USER TO-ESTABLISH AND CAN
*Textantworten formatieren.*
TEXT-RESPONSES FORMAT

b. The application can use the window to establish a dialog with the user and to format text responses.

*Die Anwendung kann das Window benutzen, um*
THE APPLICATION CAN THE WINDOW USE IN-ORDER-TO
*einen Dialog mit dem Benutzer herzustellen und*
A DIALOG WITH THE USER TO-ESTABLISH AND
*Textantworten zu formatieren.*
TEXT-RESPONSES TO FORMAT

### 2.2.2. Ing-*words*

Other highly ambiguous constructions in English involve *ing*-forms, especially in short segments. These words can be used in many ways, e.g. as nouns (gerunds), adjectives, and verbs. Due to its many functions, we will simply refer to a word occurring in this form as an "*ing*-word".

Since *ing*-words in English are one of the biggest sources of structural problems for MT, our advice is to reduce the use as much as possible. Consider the real, but truly ambiguous sentence given in (8). This sentence can be taken to mean either that the company does not take the time to improve service or that they are very quick in improving the service. The latter interpretation is likely the intended meaning, but an MT system will have difficulty determining this.

(8) At XYZ Inc. we don't waste any time improving service for our customers.

Kohl (1999) states that it is not necessary to worry about *all* occurrences of *ing*-words and specifically mentions the following cases as being acceptable:

- *ing*-words that are preceded by a preposition. A slight variation of his example is shown in (9).

  (9) For more information about printing files, see chapter 3.

  However, in the context of MT, this is ambiguous between the reading where *files* is the object of *print*, and the reading where *printing* pre-modifies *files*.
- *ing*-words that are the subject of a clause. His example is shown in (10).

  (10) Specifying the system password gives you full administrative access.

  He goes on to say: "When it's the first word of a simple sentence, an -ING can only be a gerund." In (10) this is true because there is an article (*the*) between the *ing*-word and the following noun. However, Kohl's statement is not generally true. Humans often disambiguate by applying world knowledge, but even then there may be problems as evidenced by the notorious example given in (11).

  (11) Visiting relatives can be a nuisance.

*Rule 3.* Use articles with *ing*-words when they are used as nouns; or use infinitives instead of *ing*-words, depending on what you mean.

In some cases it helps to add articles or to use the infinitive instead. Examples of this are shown in (12). (12a) shows an ambiguous use of an *ing*-word. This phrase has two major interpretations. The first interpretation can be disambiguated as either (12b) or (12c); the other as (12d).

(12) a. meeting requirements
b. meeting the requirements
c. to meet requirements
d. the meeting requirements

The sentence in (13a) is ambiguous with respect to *accommodating*, which can either modify *people* (13b) or *like* (13c). The translation into French of (13a) chose the interpretation corresponding to (13b).

(13) a. John likes accommodating people.
*John aime les personnes obligeantes.*
JOHN LIKES THE PERSONS ACCOMMODATING
'John likes people who are accommodating.'

b. John likes the accommodating people.
*John aime les personnes obligeantes.*

c. John likes to accommodate people.
*John aime accommoder les gens.*
JOHN LIKES TO-ACCOMMODATE THE PEOPLE
'John likes to accommodate people.'

In other cases it is better to avoid the *ing*-words altogether. This is true for *ing*-words that follow the object of a verb. These *ing*-words may attach to either the object or the subject.

*Rule 4.* Rewrite *ing*-words that follow an object as a relative clause or add a suitable preposition, depending on what you mean.

In (14a), taken from an IBM manual, it is unclear whether it is the *users* or the *objects* that are *using the application program*.

(14) a. Different system users may operate on different objects using the same application program.
*Les utilisateurs différents du système peuvent manipuler*
THE USERS DIFFERENT OF-THE SYSTEM CAN MANIPULATE
*les objets différents l'aide du même programme*
THE OBJECTS DIFFERENT THE-HELP OF-THE SAME PROGRAM
*applicatif.*
APPLICATION
'The different system users can operate on the different objects with the help of the same application program.'

b. Different system users may operate on different objects by using the same application program.
*Les utilisateurs différents du système peuvent manipuler*
THE USERS DIFFERENT OF-THE SYSTEM CAN MANIPULATE
*les objets différents en utilisant le même programme*
THE OBJECTS DIFFERENT IN USING THE SAME PROGRAM
*applicatif.*
APPLICATION.
'The different systems users can operate on the objects by using the same application program.'

c. Different system users may operate on different objects that use the same application program.

*Les utilisateurs différents du système peuvent manipuler*
THE USERS DIFFERENT OF-THE SYSTEM CAN MANIPULATE
*les objets différents qui utilisent le même programme*
THE OBJECTS DIFFERENT WHICH USE THE SAME PROGRAM
*applicatif.*
APPLICATION

'The different system users can operate on the different objects which use the same application program.'

The MT system that translated (14a) into French decided that it is the users who are using the same application program. If this was indeed the intended meaning, it would be safer to rewrite the sentence as in (14b), where the occurrence of *by* signals this.

However, if the intention was that the objects are using the same application program, this can easily be indicated by expanding the nonfinite clause to a finite clause as in (14c). Here the preference for close attachment makes the result come out much closer to the desired interpretation as can be seen in the translation.

*Rule 5.* Rewrite *ing*-words that are complements of other verbs.

In (15a) there is a syntactic ambiguity regarding the role of *using*, which can be either an adjunct modifier of *start* as in *you can use X to start it*, or a complement of *start* as in *start to use*. Our knowledge of motors tells us that probably the adjunct modifier interpretation is the intended one, and this can be written unambiguously as (15b) or (15c).

(15) a. The motor starts using a gas-powered pull start via a rechargeable battery.

b. You use a gas-powered pull start via a rechargeable battery in order to start the motor.

c. You start the motor with a gas-powered pull start via a rechargeable battery.

### 2.2.3. *Postnominal Modifiers*

Postnominal modifiers can cause trouble if they appear in an abbreviated form.

*Rule 6.* Do not omit relative pronouns; write *that* (*which*, *who*, etc.) explicitly.

Writing relative pronouns explicitly not only enhances MTranslatability, but also makes the sentence easier to process for the human reader. Consider how difficult to understand the sentence in (16a) is. In fact, the resulting translation into Spanish

is so bad that it is difficult to give an idiomatic translation; hence only the word-for-word gloss is given.[4] However, if you add the pronoun *that*, the sentence is easier to understand for both humans and MT systems, as shown in (16b).

(16) a. The cotton shirts are made from comes from Arizona.

*Se hacen las camisas de algodón de viene de*
IMP-REFL MAKE THE SHIRTS FROM COTTON FROM COMES FROM
*Arizona.*
ARIZONA

b. The cotton that shirts are made from comes from Arizona.

*El algodón del que se hacen camisas viene*
THE COTTON FROM-THE WHICH IMP-REFL MAKE SHIRTS COMES
*de Arizona.*
FROM ARIZONA

'The cotton one makes shirts from comes from Arizona.'

Other postnominal modifiers can also be problematic and are better expanded into full relative clauses as illustrated by (17).

*Rule 7.* Avoid post-modifying adjective phrases.

In (17a) the translation of *available* (*vorhanden*) is placed wrongly in the sentence, so that the German becomes unintelligible. This is difficult to convey in an English idiomatic translation. Translated correctly, the sentence would be as shown in (17b) where *vorhanden* is properly inflected and placed before the noun that it modifies.

The version shown in (17c) does not suffer the same mistranslation as (17a) when subjected to MT.

(17) a. The amount of adjacent space available in storage does not restrict the size of a library, or of any other object.

*Die Menge des angrenzenden Platzes vorhanden in der*
THE AMOUNT OF-THE ADJACENT SPACE AVAILABLE IN THE
*Speicherung schränkt nicht die Größe einer Bibliothek oder*
STORAGE RESTRICTS NOT THE SIZE OF-A LIBRARY OR
*irgendeiner anderen Nachricht ein.*
ANY OTHER MESSAGE IN

b. *Die Menge des in der Speicherung vorhandenen*
THE AMOUNT OF-THE IN THE STORAGE AVAILABLE
*angrenzenden Platzes...*
ADJACENT SPACE

c. The amount of adjacent space that is available in storage does not restrict the size of a library, or of any other object.

*Die Menge des angrenzenden Platzes, der in der*
THE AMOUNT OF-THE ADJACENT SPACE WHICH IN THE
*Speicherung vorhanden ist, schränkt nicht die Größe einer*
STORAGE AVAILABLE IS RESTRICTS NOT THE SIZE OF-A
*Bibliothek oder irgendeiner anderen Nachricht ein.*
LIBRARY OR ANY OTHER MESSAGE IN

'The amount of adjacent space which is available in the storage does not restrict the size of a library or of any other message.'

### 2.2.4. *Pronouns*

In many languages the pronoun has to agree in number and grammatical gender with its antecedent. This is of *some* help when translating *from* these languages, but poses serious problems when translating *into* such a language from a language like English that does not make this distinction.

*Rule 8.* Minimize use of personal pronouns.

Many commercial MT systems do not support resolution of personal pronouns, which is a rather difficult task. This means that pronouns are likely to be assigned some default number and gender, which in many cases would be misleading. Consider the well-known variation of an example of Winograd's in (18).[5]

(18) a. The police refused the students a permit because they feared violence.
b. The police refused the students a permit because they advocated violence.

These two sentences are very similar; nevertheless, the referents of *they* differ. We know that in (18a) *they* is coreferential with *the police*, whereas in (18b) *they* is coreferential with *the students*. In French, these two words have different number and gender, and this needs to be reflected in the translation of the pronoun (*they*) in order to convey the correct meaning as shown in (19).

(19) a. *La police a refusé un permis aux*
THE POLICE-FEM-SG HAS REFUSED A PERMIT TO-THE
*étudiants parce qu' elle craignait des*
STUDENTS-MASC-PL BECAUSE THAT IT-FEM-SG FEARED SOME
*actes de violence.*
ACTS OF VIOLENCE

b. *La police a refusé un permis aux*
THE POLICE-FEM-SG HAS REFUSED A PERMIT TO-THE
*étudiants parce qu' ils prônaient*
STUDENTS-MASC-PL BECAUSE THAT THEY-MASC-PL ADVOCATED
*la violence.*
THE VIOLENCE

In (19a) *elle* '*it*-fem-sg' is coreferential with *la police* '*the police*-fem-sg' and in (19b) *ils* '*they*-masc-pl' is coreferential with *les étudiants* '*the students*-masc-pl'. At present, there are no natural-language processing programs that can reliably identify the reference of pronouns. Therefore, strictly controlled languages ban the use of 3rd-person pronouns altogether.

Sometimes it is possible to avoid a pronoun just by simplifying the sentence, as illustrated in (20). The rewriting of (20a) to (20b) illustrates the following points for improving MTranslatability: (a) avoid passives, (b) shorten the sentence, and (c) avoid pronouns, all of which contribute to general clarity.

(20) a. When fasteners are removed, always reinstall them at the location from which they were removed.
b. Always reinstall fasteners in the same location.

In some cases pronouns can be avoided in this way, but since they play a crucial role in natural language, their use is a trade-off between MTranslatability and natural-sounding language. The writer (or editor) must carefully consider each occurrence on a case-by-case basis to determine whether a pronoun can be avoided. This decision may also depend on the style of the document.

### 2.2.5. *Other Rules*

*Rule 9.* Always write the complementizer *that* explicitly.

(21) a. We will do this on the condition alternatives are considered.
b. We will do this on the condition that alternatives are considered.

*Rule 10.* Avoid long noun phrases, if possible.

Long noun phrases are ambiguous with respect to the relations between the nouns. In addition, they often cannot be translated compositionally.

*Rule 11.* Always write *in order to* before an infinitive in a purpose clause instead of just *to*.

(22) a. They signed the agreement to ensure world peace.
b. They signed the agreement in order to ensure world peace.

*Rule 12.* Use one-word verbs instead of verb+particle whenever possible.

Sometimes it can be difficult to distinguish between particles and prepositions. In (23a) *up* is a particle, and in (23b) *up* is a preposition. This difference often needs to be reflected in the translation.

(23) a. She ran up a bill. (She accumulated a bill.)
b. She ran up a hill.

In addition, verb+particle (or +adverb) combinations often have many meanings, exemplified by (24a), which occurred in a mail-order catalog. This is better written as either (24b) or (24c), depending on the intended meaning.

(24) a. The shirt is designed to be worn out.
b. The shirt is designed to be worn untucked.
c. The shirt is designed to deteriorate.

### 2.3. CHECK THE STYLE

Various stylistic issues also affect MTranslatability. In this section we look at sentence length, metaphors and the like, ellipsis, passive voice, and segment independence.

*Rule 13.* Avoid overly long sentences and very short sentences.

It is well known that very long sentences can be hard to parse and hence hard to translate. But also very short segments pose problems. For example, English has a high degree of part-of-speech ambiguity; most nouns can be verbs, and vice versa. If the segment is very short, there is little context to help disambiguate.

In the segment (25a), each word can either be a noun or a verb, but the MT systems seem to prefer the multi-noun reading. This is a case where syntactic cues, described above, can be of use. A determiner added in the right place can help disambiguate the segment as shown in (25b) and (25c).

(25) a. Transfer file.
*Fichero de transferencia.*
FILE OF TRANSFER
b. Transfer the file.
*Transfiera el fichero.*
TRANSFER-IMP THE FILE
c. The transfer file.
*El fichero de transferencia*
THE FILE OF TRANSFER

*Rule 14.* Avoid metaphors, idioms, slang, and dialect.

Most MT systems have a limited coverage of metaphors, slang, and idioms. They tend to translate text at face value. An idiom like (26a) is better avoided and written as (26b).

(26) a. He got my goat.
b. He annoyed me.

*Rule 15.* Avoid ellipsis.

Ellipsis shares with pronominalization, passivization and marked word order the value of creating a cohesive, fluent text that is easier for humans to read. Used with care and thought, ellipsis is a normal and extremely useful device for facilitating communication.

Halliday and Hasan (1976) define ellipsis as omission of something in the text, with the condition that what is omitted (or *ellipted*) is *presupposed*. In the words of Halliday and Hasan (*ibid.*, 143): "An elliptical item is one which, as it were, leaves specific structural slots to be filled by elsewhere. . . . Ellipsis can be regarded as substitution by zero." Unfortunately, MT systems rarely have good sources for filling the missing slots, and ellipsis can be detrimental to MTranslatability, regardless of whether the ellipsis has a useful function in the text or not. So it is an issue worthwhile addressing in the context of writing for MTranslatabilty; it is then up to the writer to see what makes sense to do in the particular circumstances.

There are many forms of ellipsis. We have already encountered examples of ambiguity problems caused by ellipsis earlier in this section (25a) and in Section 2.2, particularly in the context of coordination. Coordination is a common environment for ellipsis to occur in – so important, in fact, that Quirk et al. (1972) place their general treatment of ellipsis in the section on coordination.

Ellipsis can be sentence-internal, as in (27). Semantics, textual context, and the emphasis tell us that the text is elliptical, with a missing verb *collects*.

(27) We collect arrowheads and our neighbor rocks.

Ellipsis can also be sentence-external, as in (28). The sentence-external case is particularly pernicious since MT systems usually translate one sentence at a time; hence, the second sentence will be likely to get an incorrect interpretation without the consideration of ellipsis.

(28) We collect *arrowheads*. Our neighbor *rocks*.

The ellipses in both (27) and (28) create part-of-speech ambiguity for *rocks*, which could be either a noun or a verb.

But often ellipsis simply creates a "malformed" sentence – malformed, at least, for an MT system that does not know how to add the presupposed parts. Consider the example in (29). In (29a), not only the article, but also the (pleonastic) subject and verb are ellipted, with disastrous results for the translation into German, whereas the full version in (29b) gives a good translation.

(29) a. Pity he won't help.
*Mitleid er wird nicht helfen.*
COMPASSION HE WILL NOT HELP

b. It is a pity that he won't help.
*Es ist schade, daß er nicht helfen wird*
IT IS UNFORTUNATE THAT HE NOT HELP WILL
'It is a pity that he will not help.'

A special kind of ellipsis is found in instructional texts, such as recipes and manuals. This style is characterized by the omission of objects that are normally obligatory, as well as omission of articles and subjects. In (30), there probably is a presupposed object of *turn*. Quirk et al. (1972) point out that some verbs have senses with obligatory complements. In the view of that work, omission of an obligatory object will *convert* the transitive sense to the intransitive sense of the same verb. For the intransitive sense the *theme* (in the sense of Jackendoff, 1972) is given by the subject, whereas in the transitive version the theme was given by the surface object. In languages like German, French, and Spanish, this difference in meaning is often accomplished by a reflexive construction for the corresponding English intransitive case.

Taken out of context, the verb in (30a) is thus an intransitive verb, with the corresponding reflexive construction in German. The sentence is ambiguous; it could contain an intransitive verb, or it could be a case of ellipsis. In the domain of instructional texts, it is likely to be ellipsis, but generally MT systems would have difficulty identifying the presupposition and resolving the ellipsis. Hence it is safer to supply an object (in the elliptical case), as shown in (30b).

(30) a. Turn every five minutes.
*Drehen Sie sich alle fünf Minuten.*
TURN YOU YOURSELF EVERY FIVE MINUTES
'Turn (yourself) around every five minutes.'

b. Turn the vegetables every five minutes.
*Drehen Sie das Gemüse alle fünf Minuten.*
TURN YOU THE VEGETABLES EVERY FIVE MINUTES
'Turn the vegetables every five minutes.'

Another example of ellipsis is given in Appendix B.1; the text contains a use of ellipsis that creates a "hurried", breathless, style.

*Rule 16.* Avoid passive constructions, if possible.

Passive voice plays a role in creating the right focus in a sentence, among other things. However, it can sometimes be hard to translate well. There are two major reasons that English passives can be difficult to translate:

– It can be very difficult to disambiguate between stative and dynamic passives, which in other languages may need to be expressed differently. This is illustrated by the Spanish translations in (31).

(31) a. The door was closed.
b. *La puerta estaba cerrada.*
THE DOOR WAS CLOSED-PASTPART (STATIVE)
c. *La puerta fué cerrada.*
THE DOOR WAS CLOSED-PASSIVE (DYNAMIC)

– The argument assignments in passive constructions may differ between source and target language, adding more complexity to the translation process.

(32) a. The company was talked about.
*Über die Firma wurde gesprochen.*
ABOUT THE COMPANY WAS TALKED
b. The horse was raced.
*Mit dem Pferd wurde ein Wettrennen gemacht.*
WITH THE HORSE WAS A RACE MADE

In (32a) there is no subject in German corresponding to the English subject *company*. This is the opposite situation of (32b) where the German has a subject, *Wettrennnen* ('race'), which does not appear in the English, and where the English subject *horse* needs to be a prepositional object in German.

*Rule 17.* Make sure that each segment can stand alone syntactically.

Since MT systems usually translate each segment as an individual unit, it is crucial for good translation quality that no segment depends syntactically on preceding or following segments. Instructional documents often structure the information in bulleted lists. Here is a danger of letting each bullet depend on the segment that leads into the list as shown in (33a).

(33) a. After you have set up your workstation, you can:
– Log on to the network
– Work locally
b. *Nachdem Sie Ihren Arbeitsplatzrechner aufgestellt haben, können*
AFTER YOU YOUR WORKSTATION SET-UP HAVE CAN
*Sie:*
YOU
'After you have set up your workstation, you can:'
– *Melden Sie sich beim Netz an*
REPORT-IMPERATIVE YOU REFL BY-THE NETWORK TO
'Log on to the network!'
– *Arbeiten Sie am Ort*
WORK-IMPERATIVE YOU IN-THE PLACE
'Work locally!'

In (33a), the two verbs in the list elements, *log on* and *work*, are dependent on *can* in the segment leading into the list. This dependency indicates that the two verbs in the list are infinitives, not imperatives. However, translated out of context, the verbs come out as imperatives, which is not unreasonable, just wrong, as shown in German translation in (33b).

The translations come out better if you make sure that the list elements do not depend on the preceding segment as shown in (34).

(34) a. After you have set up your workstation, you can do the following:
- You can log on to the network.
- You can work locally.

b. *Nachdem Sie Ihren Arbeitsplatzrechner aufgestellt haben, können*
AFTER YOU YOUR WORKSTATION SET-UP HAVE CAN
*Sie folgendes machen:*
YOU FOLLOWING MAKE
'After you have set up your workstation, you can do the following:'
- *Sie können sich beim Netz anmelden.*
YOU CAN REFL BY-THE NET REPORT
'You can log on to the network.'
- *Sie können am Ort arbeiten.*
YOU CAN IN-THE PLACE WORK
'You can work locally.'

*Rule 18.* Avoid footnotes in the middle of a segment, and make footnotes independent segments.

Some MT systems may expect footnotes to appear only at the end of a segment and hence terminate a segment when a footnote is encountered. For this reason, it is safer to place footnotes at the end of the segment and make sure that they can be treated independently.

Parentheses can pose problems similar to those of footnotes.

*Rule 19.* Do not include parenthesized expressions in a segment unless the segment is still valid syntactically when you remove the parentheses while leaving the parenthesized expressions.

## 2.4. CHECK THE PUNCTUATION

*Rule 20.* Use punctuation prudently.

Punctuation may look trivial, but it is used in various ways to give structure to a document and to indicate certain grammatical relations. One of the most important uses of punctuation is to indicate the end of one segment and beginning of another

segment. Obviously, if the MT system does not know how to segment the text correctly into individual sentences due to missing punctuation (especially periods), then the output is bound to become very peculiar. But there are other important uses of punctuation, and lack of proper punctuation can cause problems. For example, it is becoming quite common (at least in U.S. English) to omit the hyphen between a noun and a post-modifying past participle, as in (35a). This makes parsing rather difficult since *provided* may be taken as a past participle as in *the user that was provided*.[6] The parser may also take *provided* in (35a) as a past tense verb. A possible correct rewriting of (35a) is given in (35b). All in all, it is better to remember the hyphen and write the sentence as in (35b).

(35) a. *If the user provided file is not found, an error message is issued.
b. If the user-provided file is not found, an error message is issued.

Occasionally, one also encounters the opposite: wrongly placed hyphens, as in (36a).

(36) a. *He bit-off more than he can chew.
b. He bit off more than he can chew.

Commas do make a difference in intelligibility for both humans and MT systems. In (37) we show an example of what a comma can do for a sentence.

(37) a. Since Jay always jogs a mile doesn't seem that far to him.
b. Since Jay always jogs, a mile doesn't seem that far to him.

*Rule 21.* Avoid using *(s)* to indicate plural.

This construction may not translate well into other languages where subject–verb agreement may be more complicated than just a matter of an *s* in the right place. Also noun-phrase agreement, as in (38), may be difficult to handle.

(38) Notice that the pattern name must be delimited by the defined DELIMITER string(s).

*Rule 22.* Avoid using */* as in *and/or* and *user/system*.

The slash is ambiguous. It can mean either *and* or *or*, and this may affect subject–verb agreement, the rules for which may be different for the source and target languages, depending on whether *and* or *or* was intended. It can also be difficult to get noun-phrase agreement right if the gender and number differ for the two conjuncts.

Additionally, the slash is sometimes used to indicate version numbers etc. as in *System/390*, which obviously should not be given the same treatment as *user/system*.

## 2.5. CHECK THE SPELLING

*Rule 23.* Check your spelling.

If a word is misspelled, it will – at best – produce a non-translation. At worst it will prevent a successful source analysis and produce an incorrect grammatical structure.

## 2.6. CHECK THE FILE CHARACTERISTICS

This section describes some very mundane but nevertheless real problems for MT. These problems relate to overall file characteristics, which tend to be neglected by the typical user without regard for their impact. But it is important to ensure that the basic file format is in good shape.

*Rule 24.* Proofread and correct scanned documents.

Scanned documents need to be proofread and corrected since OCR software is not 100% reliable.

*Rule 25.* Avoid textual content in bitmaps.

Web pages typically have a fair amount of bitmaps. These are usually not translated by MT systems, so it is better to avoid textual content in bitmaps.

*Rule 26.* Use mark-up wisely.

Mark-up tags provide good clues for segmentation and segment type, both of which can help MT along substantially, but only if the mark-up is used in the intended way.

Use mark-up tags in a conceptual way; use header tags for headers, etc. Do not abuse tags to accomplish a purely visual effect (e.g. a header tag just to achieve a bigger font or *<br>* for a line break that does not indicate end of the segment). Use mark-up to accomplish the desired layout for tables, rather than "manual" indentation.

Specify the LANG attribute where possible. Mark any parts that are in a different language from that of the main document so that the MT system will know not to translate these parts with the same source language as the rest of the document. Write hypertext links and highlighted text such that they can be translated as a single entity. This way the mark-up will look better for the translation. Mark strings that should not be translated.

Make sure that words that are used as labels or names are properly identified.

Use ISO 8859 or Unicode characters throughout, and, ideally, use character entities for characters that are not part of the 7-bit ASCII character set. For instance,

in HTML source code, *u*-umlaut (*ü*) should be represented by the `&uuml;` character entity.

## 3. Tools for Improving MTranslatability

In this section we give an overview of the main types of authoring tools that are available for helping a writer improve MTranslatability.

For each type of tool, we give a brief description of what its intended function is and evaluate the usefulness of that function in the context of MTranslatability. In addition, we give a brief description of some specific systems.

In Section 3.1 we look at spell checkers, in Section 3.2 we investigate the use of grammar and style checkers, and in Section 3.3 we treat CL checkers. Section 3.4 describes annotation tools and tagging. Finally, in Section 3.5, we discuss the use and limitations of the built-in tools for detecting unknown words that often come with MT systems.

### 3.1. SPELL CHECKERS

The objective of spell checkers is to point out misspelled words and, where possible, to suggest the correct spelling. Most spell checkers work with a dictionary. If a word is not found in the dictionary (including user-defined dictionaries), it will be flagged as a misspelling, and alternatives will be given. In Section 2.5 we mentioned the importance of checking the spelling before attempting MT; however, spell checkers are limited in their functionality in that they do not generally discover words that happen to be valid words, but incorrect in context. For example, there is no way an ordinary spell checker can diagnose *there* as a misspelling for *they're* in *There very happy.* For some funny examples of what this kind of misspelling can lead to, see Zar (1994) and Anonymous (1995).[7]

### 3.2. GRAMMAR AND STYLE CHECKERS

One objective of grammar and style checkers is to point out ungrammatical constructions. This is a very difficult process because the checkers have to try to make grammatical sense out of potential grammatical nonsense. Since the precision is low, writers often dislike using them. Nevertheless, grammatical input to MT stands a better chance of getting a good translation. It is, however, not *sufficient* to guarantee a correct translation. Another objective is to point out complexity, repetition, certain syntactic structures, etc., in order to help the writer produce more stylistically pleasing text.

For English, we looked at the grammar and style checkers that come with some common text editors: Microsoft Word2000, Word Pro 97 (CorrecText Grammar Correction System), and Corel WordPerfect, version 7 (Grammatik).

For German, we looked at MULTILINT's German grammar and style checker (Schmidt-Wigger, 1998; Reuther, 1998; Reuther and Schmidt-Wigger, 2000; Multilint, nd) and FLAG German grammar checker (Becker et al., forthcoming; Bredenkamp et al., 2000; DFKI, 2000).

We categorized each check into one of three categories: (a) Useful for MTranslatability, (b) not useful for MTranslatability, and (c) more or less harmful for MTranslatability. The grammar checkers show a tendency to lump together different types of problems into one type of check. Some of the problems are more relevant for MTranslatability than others; hence, some checks (representing more than one type of problem) belong to more than one usefulness category, depending on which aspect you are looking at.

Not surprisingly, the vast majority of checks fall into the "useful" category because they directly address issues such as spelling errors and ungrammatical constructions. Let us just mention a few of the useful checks here: punctuation errors, long noun strings, possessives versus plurals, and subject–verb agreement.

However, there were also some recommendations that are directly opposed to MTranslatability. These recommendations included sentence variety and avoidance of certain contractions like *I'm* and *they've*, which may actually help the parser as opposed to contractions like *it's* and *we'd*. Certain contractions reduce ambiguity, while others increase ambiguity, so they have to be treated separately in the context of MTranslatability. The phrase in (39a) must be interpreted as (39b) due to the contraction; the interpretation shown in (39c) is not possible.

(39) a. The books I've read well.
b. [The books [that I have read well]].
c. *[The books [that I have]] read well.

In between these two categories there are also a few checks that are neutral with respect to MTranslatability. These include checks for gender-specific words, one-sentence paragraphs, and use of first person.

Overall, grammar and style checkers demonstrate a limited usefulness in the preparation of a document for MTranslatability. As long as the user is aware of the fact that some recommendations are at odds with MTranslatability, these checkers can be considered helpful tools. However, they are not *sufficient* since they do not address issues of ambiguity, which is a serious drawback.

### 3.3. CONTROLLED LANGUAGE CHECKERS

A CL is a form of language with special restrictions on grammar, style, and vocabulary usage. The objective of a CL is to improve consistency, readability, translatability, and retrievability. This is achieved by putting constraints on what the writer can say, so as to reduce ambiguity and complexity.

CLs can be divided into two major categories: languages whose intended audience is non-native speakers, and languages whose main "audience" is MT. These

two categories have characteristics in common, but also differ in certain respects. For a further overview, the reader is referred to Huijsen (1998), Wojcik and Hoard (1997), Mitamura and Nyberg (2000).

### 3.3.1. *KANT Controlled English*

KANT Controlled English from Carnegie Mellon University (Mitamura and Nyberg, 1995; Nyberg and Mitamura, 1996; Mitamura, 1999) was designed with MT in mind. This CL aims at balancing the control of the vocabulary with the control of the grammar. In this way, the writer is not forced to write very convoluted sentences in order to stay within the controlled vocabulary.

The constraints of KANT Controlled English are divided into the categories of vocabulary constraints, phrase-level constraints, and sentence-level constraints.

The vocabulary constraints include restrictions on the use of *ing*- and *ed*-words, pronouns, and conjunctions; limitation for each word to a single meaning per part of speech; clear specification of the sense and use of modal verbs; clear specification of approved orthography and use of abbreviations; and advocacy of the use of determiners.

The phrase-level constraints include: avoidance of verbs with particles; repetition of prepositions in coordinated prepositional phrases; and prohibition of coordination of verb phrases.

The sentence-level constraints require parallelism in coordination; explicit relative pronouns; and avoidance of ellipsis.

This list clearly addresses a number of the issues that we described earlier in Section 2, and consequently all of these restrictions enhance MTranslatability. This is not surprising given that this CL was designed with the express purpose of improving MTranslatability.

The KANT technology for CL and MT has been successfully deployed in document production for Caterpillar Inc. (Kamprath et al., 1998).

### 3.3.2. *AECMA Simplified English*

AECMA Simplied English (SE) (AECMA, 1995) is the aerospace industry's CL. It is used for making aircraft maintenance manuals unambiguous and easier to read for non-native speakers of English. This is useful since aircraft have to be maintained by local mechanics in airports all over the world.

*The MAXit Checker.* The MAXit AECMA SE Checker (Smart, 1998) offers a number of useful checks, including checks for verbs with particles, wrong punctuation, long sentences ($>$ 21 words), gerunds, and lack of parallelism.

There are also a few AECMA-specific checks that are not useful in the context of MTranslatability, e.g. gender-specific pronouns and "safety warning required". This is not surprising, given that AECMA Simplified English was not designed with MT in mind.

*Boeing Technology.* Boeing offers a good AECMA SE Checker (Wojcik et al., 1990; Hoard et al., 1992; Wojcik and Holmback, 1996). In addition to checking for SE compliance, the Boeing SE Checker also catches mistakes like lack of subject–verb agreement, repeated words, misspelled words, and punctuation problems. Boeing has also developed a word-sense disambiguator, which is described in Holmback et al. (2000). It is used for checking compliance with AECMA SE word meanings.

The Boeing Technical English (BTE) Checker (Wojcik et al., 1998) is a modified version of the Boeing SE Checker that supports more general technical writing. The intended audience of BTE is also non-native speakers of English. BTE differs from AECMA SE in two regards: the writing rules and the controlled vocabulary. On the one hand, the writing rules have been relaxed, e.g. the blanket ban on *have* and *be* auxiliaries has been relaxed to a recommendation; on the other hand, new rules have been introduced, e.g. requirement of subordinating *that* and of relative pronouns. Addition of these rules appears to support the overall idea of MTranslatability.

AECMA SE has a very limited vocabulary, which has been expanded for BTE. This has given rise to more word-sense ambiguities, which would clearly be an issue in the context of MT. However, the setup of the above-mentioned word-sense disambiguator is general enough to be used with BTE, provided the disambiguator's dictionaries are extended.

### 3.3.3. *Controlled Automotive Service Language*

GM's Controlled Automotive Service Language (CASL) English described in Means and Godden (1996) has 62 rules that were designed for the express purpose of increasing the quality of MT output for vehicle service information.

Godden (2000) describes the various stages of design that CASL has gone through, from author-centered, through editor-centered, to a hybrid model. The hybrid model puts the final responsibility on the authors rather than on the CASL editors, but only requires the authors to adhere to the eight most important rules of the total rule set. The CASL editors then take care of the remaining rules.

The eight most important rules are the following: do not write sentences more than 25 words long; do not use noun clusters longer than four words; do not use pronouns, except *you*; use articles before nouns; use *that* to clarify sentence structure; repeat the head noun with conjoined adjectives; use only CASL-approved terms; and do not use parentheses for explanations.

All these rules will obviously help MTranslatability. And of course it may be undesirable for writers of ordinary text (non-CASL documents) to avoid pronouns completely.

A very interesting property that sets CASL apart from other CL checkers is the special CASL tag (Godden, 1998b). This SGML tag indicates compliance or non-compliance with the CASL rules. When the MT system detects CASL compliance,

it knows which interpretation of an ambiguous construction is the approved and intended one.

Great care has also been taken to reengineer the entire business process to take advantage of CASL (Godden, 1998a) and to educate the writers (Means et al., 2000). These are areas that are often overlooked when a company attempts to introduce the combination of CL and MT. Including these aspects no doubt contributed significantly to the success of CASL.

### 3.3.4. *MULTIDOC*

Components of MULTILINT are now part of a more comprehensive technology for controlled authoring. It covers German, English, French, and Swedish. Originally developed for the European automotive industry, it is now also used in other technical production domains, including Siemens, Germany, and Sun Microsystems (Schütz, 2001).

It has evolved into a full-fledged open-architecture technology that includes a spell checker, a term checker, a style checker and a grammar checker, thus straddling the border between spell and grammar checker on the one hand and CL authoring tool on the other. According to Haller (2000), the goal of the source-text modules is to ensure a more readable, understandable and translatable text. Not surprisingly, Haller cites "pure gerunds"[8] in English as a source of misunderstanding that the checker should flag. One of Haller's examples involves the infamous construction of *ing*-word following the object of a verb (40):

(40) Remove the setscrews securing the bridging plate.

### 3.3.5. *EasyEnglishAnalyzer*

IBM's EasyEnglishAnalyzer (EEA) tool (Bernth 1997, 1998, 1999c) is an authoring tool that points out ambiguity and complexity, thereby helping writers produce documents that are more MTranslatable. EEA also does some standard grammar checking. EEA is used by information developers in IBM with the aim of improving both readability for an international audience and to prepare selected documents for MT, depending on which checks are set on.

The checks that are useful for MTranslatability include checks for nonfinite verb phrases, ambiguous scope in coordination, long sentences, incomplete sentences, and non-parallelism, as well as more traditional grammar checking. In addition to these, some checks that are not directly aimed at improving MTranslatability have been included in order to accommodate corporate writing guidelines. These include checks for abbreviations and restricted or prohibited words.

EEA's Clarity Index summarizes the problems that are encountered in a given document as a single number that indicates the clarity for the whole document. In this context, the Clarity Index can also be viewed as an indication of MTranslatability. The problems are weighted according to severity (impact), context, and

document size. EEA also includes ETerms, which collects multinouns and unknown words along with their frequency. These are candidates for terminology to be added to the user lexicons.

### 3.4. ANNOTATION

A very different way to prepare a document for better MTranslatability is annotating (or tagging) it. This method is used for various purposes, such as mark-up for formatting purposes or for enriching the semantic and knowledge content of documents. It is also used for easier accessing and processing of information on the World Wide Web. Two workshops were held following the Coling conference in August of 2000 – one on syntactic annotation and one on semantic annotation. Both workshops included presentations and discussions on tools and techniques for linguistic annotation. In Sections 3.4.1 and 3.4.2 we give a brief description of two Japanese annotation efforts, Global Document Annotation and Linguistic Annotation Language. The European OTEXT standard is mentioned in Section 3.4.3.

#### 3.4.1. *Global Document Annotation*

Global Document Annotation (GDA) (Hasida 1997, 2000) is an extension of XML. The aim is to have authors of web pages annotate their documents with semantic and syntactic tags so as to help not only MT, but also other text-understanding systems.

In (41) we show an example of GDA annotation.

(41) Time flies like an arrow.

```
<su>
<np sem = time0> time</np>
<v sem = fly1>flies</v>
<adp>like<np>an arrow</np></adp>
</su>
```

The XML elements such as <np>... </np> encode parse-tree bracketing, and the property `sem` disambiguates polysemy of words. The word senses in this particular example (`time0` and `fly1`) are based on WordNet senses, but the plan is that a growing population of GDA users will develop their own ontologies for all languages.

The way that such an XML tagger improves MTranslatability – assuming all MT engines are modified to recognize the tags – is obvious: Some of the hardest problems for the MT parser will be solved. Ambiguities on both the syntactic and the semantic levels will be resolved, and proper nouns will be identified.

The inventory of GDA tags is very comprehensive. In addition to syntactic and semantic word disambiguation, it includes tags for scoping, tense and aspect, indicators of levels of politeness, and types of utterances. Consequently, it is enormous.

Without an efficient and user-friendly interface, using the tags seems a daunting task. But doubtlessly, if the tags are used and MT engines can interpret them, the translation output will improve dramatically. An interactive editor for GDA has been developed.

### 3.4.2. *Linguistic Annotation Language*

Linguistic Annotation Language (LAL) (Watanabe et al., 2000), another XML-compliant tag set, consists of linguistic information tags and of task-dependent instruction tags.

*Linguistic information tags* include both syntactic and semantic tags.

**Syntactic tags** These include tags that identify sentence boundaries; tags that denote word information (including attributes such as base form, semantic type, unique word ID, part of speech, dependencies, and language-specific features such as number, gender, tense etc); and tags that denote phrase and clause boundaries. Besides boundaries, dependencies can also be expressed by using IDs and modifier attributes of the word tag.

**Semantic tags** These are user-definable tags that include tags indicating proper names (of persons, places, organizations etc.), acronyms (and other abbreviations), dates, times, numbers, and monetary units.

*Task-dependent instruction tags* include a tag that indicates whether a piece of text should be translated or not, and a tag that indicates whether a piece of text should be considered for summarization purposes.

LAL tags are usually expressed by using XML namespaces. Their XML namespace prefix is `lal`. LAL tags interact with two types of programs: NLP systems for generating and using the LAL annotation, and an annotation editor.

The Slot Grammar parser (McCord 1980, 1990) used in IBM's MT systems for source English, German, French, Italian and Spanish generates and accepts LAL annotation. For Japanese, a post-processing routine converts the output of the Japanese KNP parser (Kurohasi and Nagao 1994, 1998) into LAL format. The annotations produced in these ways are used as input to the annotation editors for English and Japanese. This means that ambiguities can be resolved by using the annotation editor to pre-edit the source text before translation into several languages.

The *annotation editor* allows the user to edit the LAL annotation of a text. This editor is interfaced to the LAL-generating grammar, which provides candidate annotation for each segment. A human editor can then use the annotation editor's graphical user interface to check over the automatically produced annotation and change it as necessary. The user can do this without having to see the tags by working on the graphical representation of the tree; the changes are then reflected in the internal LAL annotation. LAL annotation is distinguished from previous tag-

defining efforts by providing a comprehensive, yet simple list of annotation tags. Keeping things simple is crucial for user acceptance.

Examples of LAL-annotation are given in (42).

In (42a) *IBM* is marked as an acronym with expansion 'International Business Machines'. Example (42b) illustrates pronoun resolution. Note that not *all* words need to be annotated.

(42)a. IBM

```
<lal:acronym expan="International Business Machines">
  IBM
</lal:acronym>
```

b. The cat chased a mouse. After it caught it, it ate it.

```
<lal:s>
  The
  <lal:w id="w1">cat</lal:w>
  chased
  <lal:w id="w2">a mouse</lal:w>.
</lal:s>
<lal:s>
  After
  <lal:w ref="w1">it</lal:w>
  caught
  <lal:w ref=w2>it</lal:w>,
  <lal:w ref="w1">it</lal:w>
  ate
  <lal:w ref="w2">it.</lal:w>
</lal:s>
```

In (42b), each sentence is enclosed in the `<lal:s>` ('sentence') and `</lal:s>` tags. Unique, cross-sentential IDs are assigned to *cat* (`id=w1`) and *mouse* (`id=w2`). The `ref` value is used for tying in the connections between the pronouns and their antecedents, and the human editor can mark the `ref` value appropriately, using the values supplied in the `id` fields.

### 3.4.3. *OTEXT*

OTEXT (Thurmair 1997, 2000a) is a subsystem of the OTELO project, a collaborative effort between the European Union and a consortium of industrial partners. The objective is to design and develop a comprehensive automated translator's environment. The project partners have developed a standard set of tags for exchanging documents across different MT and translation memory systems.

Some MT-specific tags mark strings that should not be translated by the MT system. The `<pr>` ('protect') tag protects strings that are not part of the text flow. These are typically parameter settings, internal control information etc. The `<l>`

('literal') tag protects strings that are part of the text flow such as a piece of code, or an address. In contrast to these two tags, the <sp/> ('special character') tag specifies characters which have special meanings and that need to be preserved by the MT system. Examples are soft returns, hard blanks etc. Finally, the <tu> ('text unit') tag is used to indicate segmentation.

### 3.5. USER DICTIONARIES AND AUTOMATIC DETECTION OF LEXICAL INADEQUACIES

It is important to build user dictionaries. Most MT systems supply a utility that enables the user to detect words or phrases that are not listed in the dictionary. But not all unknown words will be identified. This can happen if a word or a phrase appears in the dictionary with some, but not all, possible parts of speech. If the applicable part of speech is missing, it will obviously not translate well, and at the same time it will not appear in the list of "unfound" or "unknown" words. For example, the phrase *OK* may be listed in the dictionary as an adjective only. It is in the dictionary and so will not be flagged as an unknown word. If the document, however, uses the word as a verb, and this is not covered in the dictionary, the translation will suffer accordingly.

Another possible shortcoming is that a word does appear in the dictionary but not in connection with the semantic sense and appropriate transfer that is required for the document to be translated. For instance, the English word *pig* may be in the lexicon as referring to an animal, and hence the German transfer *Schwein*. However, if the document to be translated deals with the domain of oil production, where a *pig* refers to a technical device, it should be translated as *Molch* ('newt').

Because of such deficiencies of a simple, context-free, dictionary look-up, some MT systems come with more context-sensitive listings that allow querying the coverage for a particular domain or subject area, or generation of a list of all content words with their anticipated translation in context. Checking such a list is time-consuming, but rewarding, if one finds uncovered entries or transfers.

The dictionary also needs to contain noun strings that cannot be translated compositionally but have to be treated as a unit. Terminology-collection tools are useful for gathering candidates for these entries. These tools typically work by gathering noun strings and sorting them by frequency. High-frequency noun strings are likely to be special terminology that cannot be translated compositionally.

## 4. Ways to Measure MTranslatability

In this section we describe various approaches to automatic assessment of the MTranslatability of a document. In Section 4.1 we argue that readability scores do not have anything to offer in the context of MTranslatability, and in Section 4.2 we give a brief description of two tools that are useful for assessing MTranslatability.

### 4.1. AUTOMATIC READABILITY SCORING

Automatic readability scoring is often provided with standard grammar checkers such as those packaged with Microsoft Word2000, Lotus Word Pro 97, and WordPerfect. These scores are designed for human readability, not MTranslatability, and are based on sentence length and word length. Shorter words and shorter segments are considered easier to read for humans. But shorter words are often more ambiguous and therefore more difficult to translate well by an MT system. And very short segments (four words or less) are very ambiguous in English due to the great ambiguity of part of speech in English.

We built a short test corpus of problematic sentences and edited them according to the recommendations in Section 2. We found that the corpus showed improved clarity and translatability after pre-editing, but at the same time it achieved a reduced readability score. One would assume – and many writers claim it – that readability and translatability are almost synonymous, or at least that one is a prerequisite of the other. It turns out that this is not the case, at least not with the automated readability scores incorporated with the common word processors. Shehadeh and Strother (1994) report on a survey on computerized readability formulas that the authors undertook. The paper criticizes existing readability scales for not taking into account such factors as organization, clarity, syntax and structure.

### 4.2. AUTOMATIC MTRANSLATABILITY SCORING

In this section we will give a brief description of two different ways of measuring MTranslatability.[9] The Logos Translatability Index (LTI) gives an overall indication of the MTranslatability of a document as a whole, whereas the IBM Translation Confidence Index (TCI) gives a rating of its confidence in the translation by a given MT system for each sentence. Viewed in a certain way, the TCI is also an indication of the MTranslatability of a document.

#### 4.2.1. *The Logos Translatability Index*

In the early 1990s, researchers at Logos Corporation developed a utility prototype (Gdaniec, 1994) that automatically measures and scores the suitability of English and German documents for the Logos MT system.

The LTI is based on gross statistical properties of a document rather than on parsing the sentences. This was suggested by the fact that there appeared to be a rough correlation between the quality of raw MT output and certain gross properties of the text, such as length of the sentences, degree of syntactic complexity, discourse characteristics, etc. Although the LTI score is derived on the basis of gross sentence properties, sentence-specific information cannot be provided with any degree of reliability because there is no full-scale parsing.

The program starts off with a score of 7 and then penalizes the sentences for negative properties. The decision as to the minimum score that a document must

reach in order to be acceptable for gisting or post-editing purposes is subjective. There is no absolute, objective threshold.

Negative sentence properties are: too long or too short; words not found in the MT dictionary; short parentheses; coordination; homographs; interrogatives; unmatched parentheses; embedded clauses; part-of-speech ambiguities; certain ambiguous words (such as *ing*-words, *as*, *with*, etc.), and so forth.

Before translation, the user can have the document scored by the LTI program. It will return with a score and a recommendation such as "This document is not suitable for MT" or "This document is conditionally suitable for MT". The LTI would also suggest why a particular document is not or only conditionally suitable. It would tell the user, for instance: "The sentences on the whole are too long; Sentence # N is far too long; The document contains many words and compounds that are not in the dictionary. Run your document through the New-Word-Search utility and update your dictionary; The document contains many difficult words such as . . . "

The user can make changes in the document in order to decrease complexity and ambiguity, and can update new words and phrases. Thus, the LTI can provide users with a measure that not only correlates with the quality of the MT output, but also helps them modify their source document in such a way as to improve the MTranslatability.

### 4.2.2. *The Translation Confidence Index*

IBM's Translation Confidence Index (TCI) (Bernth, 1999a; Bernth and McCord, 2000) automatically provides an index of the MT system's own confidence in its translation, for a given segment. In other words, the TCI returns a translation quality value for each segment. This value can be used to mark segments that need special attention during post-editing. The confidence value is calculated during the various stages of the MT process. It is based on such factors as parse scores, ambiguity, difficult constructions, lexical coverage, and success of structural generation (transformations) (Bernth and Gdaniec, 2000). These factors can be set on or off in the TCI's language-pair-specific user profile. Whereas the TCI was designed to give an overall picture of the expected quality of the MT output by taking all aspects of the MT process into account, the parts that deal with source analysis give a picture of the general MTranslatability. Hence, turning all non-source-language-specific factors off in the user profile in effect gives an Mtranslatability score that can be independent of the target language. On the other hand, the TCI score will give the translatability for a particular language pair for a specific IBM MT system when *all* aspects are taken into consideration.

## 5. Conclusion

In this section we summarize our findings about what the writer can do and what MT researchers need to resolve. In addition, we address the issue of whether it is worth the effort to follow the recommendations in this paper.

### 5.1. WHAT THE WRITER CAN DO

As we have shown in this paper, there is actually a lot that the writer can do to improve MTranslatability. Most of the things are not difficult, but require some awareness of linguistic issues. It is our hope that this checklist, which summarizes our findings, can be of use to the writers:

- Write grammatically
- Write unambiguously
- Consider style
- Punctuate correctly
- Spell correctly
- Use mark-up correctly
- Use syntactic and semantic annotation

Of course, not every writer will want to do *everything* we recommend in this paper, but it seems reasonable to expect MT output quality to increase proportionally to the degree that the recommendations can be followed. An important thing to remember in the context of MTranslatability is that what makes life easier for the human reader is not always useful in the context of MT.

### 5.2. WHAT MT RESEARCHERS NEED TO SOLVE

There are at least two levels of work for the MT developers and researchers. One immediate level is to enable MT systems to take advantage of mark-up and syntactic and semantic annotation. In order for this to be realistic, it is also necessary to settle on some standards for annotation. This is likely to be a two-way process, in that the MT developers will have to know which schemes are popular (i.e. worthwhile recognizing), while the annotation schemes that can be handled by commercial MT systems probably also will be the ones that people will use. Thus a feedback loop is created.

Another immediate item would be for MT systems to be able to recognize ambiguity and supply the user with more than one translation to choose from. In many cases it would probably be much more obvious to the user than to the MT system what the intended meaning was.

The checklist given in 5.1 should be seen as an intermediate measure until MT systems get better so that they are able to deal with these issues. The main problems seem to stem from lack of context, lack of world knowledge, and lack of ways to use these to resolve ambiguities, and from authors' mistakes. In the longer term, these issues need to be addressed by the MT researchers. These are difficult issues,

which we should not expect to solve overnight. One approach is to model human world knowledge and reasoning.

### 5.3. IS IT WORTH THE TROUBLE?

In this paper we have given many recommendations as to what a writer can do to improve MTranslatability, and one might well ask the questions: "Is it worth the trouble? Will it help?" before embarking on extensive editing or on writing what might be perceived as unnatural English.

Unfortunately we do not have the resources to conduct an *extensive* test of our claims. However, we can supply three arguments in support of our claims.

First of all, we have provided an analysis of *why* these phenomena can cause an MT system trouble, and have given examples of unacceptable output. For example, it is clear that structural ambiguity – be it *real* structural ambiguity or just *accidental* structural ambiguity (Hutchins and Somers, 1992: 88ff) – poses challenges to most current MT systems.[10] Likewise, *omission* of information, be it ellipsis or missing sentence delimiters, causes the MT system to go into what we might call "guessing mode".

Second, we refer the reader to the following two studies related to these issues.[11] Bernth (1999b) describes a small study of the improvements of output quality for the LMT MT system (McCord and Bernth, 1998) when applying the EasyEnglishAnalyzer rules. "Useful" translations (as judged by a native speaker of the target language) were about 68% without taking EasyEnglishAnalyzer's recommendation into consideration, but soared to 93% with pre-editing for the given text sample (technical document).

Godden (2002) reports on a study done at GM to evaluate the effect on MT of applying CASL rules to an existing English automotive maintenance text. The text was rewritten to conform to 30 rules, and the bilingual dictionaries updated to cover the text. With these preparations done, the original and the pre-edited texts were machine-translated into French. The results were then rated by two persons from GM Canada, one of them a certified English⇔French translator, the other an expert bilingual automotive technician. The ratings covered the following categories: "correct", "partially correct", and "wrong". The exact numbers are confidential, but Godden reports a *very* significant increase in percentage of correct translations for the pre-edited version over the original version, as well as a *very* significant decrease in percentage of wrong translations. The results clearly showed that CL pushed some sentences from the "wrong" category to the "partially correct" category, and others from that category to the "correct" category. Godden reports on additional results, namely savings in post-editing for a different corpus. This was a pilot study involving a service manual that was first rewritten into CASL English and then machine-translated into French. The MT output was outsourced for post-editing, and the final French was published in Canada. The goal was to reduce the cost of translation by at least 80%; however, as it turned out, GM beat their

cost targets. This was the validation of the business case, and showed that CL can contribute in a positive business sense.

Third, we have conducted a test on two small samples, one with English source, and one with German source. We edited the samples according to our recommendations, and translated the unedited and edited versions.

The English sample (APS, 1997) was translated into German, French, and Spanish, using the same MT systems that supplied the translations in our one-sentence examples. For each target language, we used *one* of the MT systems, but a different system for each. Obviously, the results would have been less system-dependent and hence more reliable if we had used several MT systems for translating into each target language and averaged the results.

The sample comprised 69 sentences in the domain of plant care instructions, and the text type was Q&A. Of these, we edited 44 sentences. Of the 44 edited sentences, 43 sentences caused a difference in output for MT system 1 (French), 37 for system 2 (German), and 34 for system 3 (Spanish).

Out of the total set of rules given in this article, the following rules were relevant and hence applied: 1, 2, 3, 4, 7, 9, 11, 13, 14, 15, 19, 20, and 23. Rules 15, 13, 20, and 11 were used the most. For 12 sentences, more than one rule was applied. Rules 11, 13, 15, and 20 were often involved in these rule combinations. Appendix C gives one paragraph of the document, in its original and revised forms, with the corresponding translations into French.

The resulting output was evaluated by three native speakers of each language, who also know English well. These evaluators had access to both source and target texts. The three evaluations were averaged for each language. The scores are summarized in Table I. The scale goes from 0 (worst) to 5 (best).[12] Of course there were variations in the scorings for each language, but the evaluators did consistently agree that the revised documents showed improvements. Interestingly, the evaluators who scored the translations of the unrevised document the lowest showed the highest improvement for the translations of the revised version.

The German sample (Hydroplant AG, 2001) consisted of 14 sentences. We translated it into English with one system and asked three native English speakers to score the result. The evaluators did not have access to the source text. The average scores are shown in Table I. The sample had to be substantially edited because of non-standard spelling and punctuation. Therefore it does not seem fruitful to make a distinction between the scores for the whole document and the rewritten parts. The German⇒English texts are shown in their totality in Appendix B.

Regarding the English⇒FGS evaluations, the translation quality went up 4–15% for the total document and 25–36% for the edited sentences, depending on language pair and MT system, as indicated by the numbers in Table I. According to the interpretation of the scoring scheme used, the edited sentences were shifted from around "compromised intelligibility" (3) to around "mostly intelligible" (4).

Adherence to our rules has the biggest impact if *lexical* problems do not interfere with the intelligibility of the translation. The impact is reduced if a bad

*Table I.* Summary of Evaluation Scores (E = English, F = French, G = German, S = Spanish)

| MT System<br>Language | MT1<br>E ⇒ F | MT2<br>E ⇒ G | MT3<br>E ⇒ S | MT4<br>G ⇒ E |
|---|---|---|---|---|
| Total document, unedited | 3.31 | 3.75 | 3.54 | 3.67 |
| Total document, edited | 3.44 | 4.30 | 3.75 | 4.59 |
| Edited sentences, before change | 2.98 | 3.19 | 3.15 | n/a |
| Edited sentences, after change | 3.78 | 4.35 | 3.95 | n/a |

lexical transfer renders the translation of a syntactically and grammatically correct sentence unintelligible. This fact is shown in the consistently and proportionately higher improvement rates for MT system 2 (36%), which could be calibrated for the domain-appropriate terminology.[13]

One thing to bear in mind when considering whether writing for MTranslatability is worthwhile, is that it is much easier to follow our recommendations *from the start* than to make changes once the text has been written with no view to MT. If the text is not written with MT in mind, an additional consideration is also how much the editor can and wants to interfere with the original text. The amount of editing necessary to obtain the results reported on in our evaluation is illustrated by the examples in the appendices.

## 6. Acknowledgements

We extend our sincerest thanks to Esmé Manandise and Michael McCord for general comments on the paper; to Francisco Barahona, François D'Heurle, Sylvie Levesque, Consuelo Rodríguez, Esmé Manandise, and Gabriel Silberman for native-speaker advice on some of the non-English examples; and to the twelve evaluators for scoring the translation output used for the study described in Section 5.3. We also thank the reviewers of the paper for their helpful comments and suggestions.

## Appendix

### A. Summary of Rules

1. Avoid ungrammatical constructions.
2. Repeat final words of the left conjunct or initial words of the right conjunct, as necessary, to disambiguate the coordination.
3. Use articles with *ing*-words when they are used as nouns; or use infinitives instead of *ing*-words, depending on what you mean.

4. Rewrite *ing*-words that follow an object as a relative clause or add a suitable preposition, depending on what you mean.
5. complements of other verbs.
6. Do not omit relative pronouns; write *that* (*which*, *who*, etc.) explicitly.
7. Avoid post-modifying adjective phrases.
8. Minimize use of personal pronouns.
9. Always write the complementizer *that* explicitly.
10. Avoid long noun phrases, if possible.
11. Always write *in order to* before an infinitive in a purpose clause instead of just *to*.
12. Use one-word verbs instead of verb + particle whenever possible.
13. Avoid overly long sentences and very short sentences.
14. Avoid metaphors, idioms, slang, and dialect.
15. Avoid ellipsis.
16. Avoid passive constructions, if possible.
17. Make sure that each segment can stand alone syntactically.
18. Avoid footnotes in the middle of a segment, and make footnotes independent segments.
19. Do not include parenthesized expressions in a segment unless the segment is still valid syntactically when you remove the parentheses while leaving the parenthesised expressions.
20. Use punctuation prudently.
21. Avoid using *(s)* to indicate plural.
22. Avoid using */* as in *and/or* and *user/system*.
23. Check your spelling.
24. Proofread and correct scanned documents.
25. Avoid textual content in bitmaps.
26. Use mark-up wisely.

## B. German to English translation

### B.1. ORIGINAL INPUT

*hallo pflanzendoktor, habe mir vor wenigen monaten einen bananenpflanze und eine fächerpalme(?) gekauft. zusammen mit 2 weiteren palmen (die eine ist eine dracaena marginata). die blätter der bananenpflanze und der fächerpalme werden nun aber seit einiger zeit immer trockener und sterben ab, die pflanzen sind aber ausreichend feucht. liegt das an der luftfeuchtigkeit und wenn ja: wie kann ich die erhöhen und dem entgegenwirken? vielen dank!*
*Der Pflanzendoktor:*
*Mangelende Luftfeuchtigkeit kann die Ursache der braunen Blätter und Ränder sein. Wenn die Blätter abtrocknen, hat das wahrscheinlich andere Gründe. Spinnmilben wäre eine Möglichkeit, Bananen sind sehr anfällig auf Spinnmilben. Auf der*

*Blattunterseite müssten winzige Tierchen erkennbar sein und feine Spinnweben in den Blattachsen. Die Blätter wirken durch das Saugen des Schädlings farblos und matt. Hohlen Sie in einem Fachhandel ein Mittel gegen Spinnmilben, behandeln Sie die Pflanze mit Wiederholungen. Wichtig: vor der Behandlung, so viele Blätter wie möglich abschneiden.*

### B.2. TRANSLATION OF ORIGINAL TEXT

*Hello plant doctor, bought a banana plant and one for fan palm (for ?) a few months ago for me. Together with 2 additional palm trees (the/that one is a dracaena marginata.) The/that scrolls becomes of the banana plant's and of the fan palm, however, for some zeit more dry and drier now and dies, planting these are, however, sufficiently damp. Does that lie, at the humidity of the air and if so: How can I increase these and counteract this? Many thank!*
*The plant doctor:*
*Mangelende humidity of the air can be the reason for the brown leaves and edges. If the leaves dry up, that has other reasons probably. Spider mites would a possibility be, bananas are very susceptible to spider mites. On the underside of the leaf(s), minute small animals had to be recognizable and fine cobwebs in the leaf axles. The leaves appear colorless and dull through the sucking of the pest. Cavernous you treat a means at a specialized dealer against spider mites you the plant with repetitions. Important: Cutting off as many leaves before the treatment as possible.*

### B.3. EDITED INPUT

The text has been rewritten to correct typos; to provide upper case for nouns; to cut run-on sentences into two with proper punctuation; to add omitted subjects; and to leave out superfluous commas.
*Hallo Pflanzendoktor. Ich habe mir vor wenigen Monaten eine Bananenpflanze und eine Fächerpalme gekauft. Zusammen mit 2 weiteren Palmen (die eine ist eine dracaena marginata). Die Blätter der Bananenpflanze und der Fächerpalme werden nun aber seit einiger Zeit immer trockener und sterben ab. Die Pflanzen sind aber ausreichend feucht. Liegt das an der Luftfeuchtigkeit? Wenn ja: wie kann ich die erhöhen und dem entgegenwirken? Vielen Dank!*
*Der Pflanzendoktor:*
*Mangelnde Luftfeuchtigkeit kann die Ursache der braunen Blätter und Ränder sein. Wenn die Blätter abtrocknen, hat das wahrscheinlich andere Gründe. Spinnmilben wäre eine Möglichkeit. Bananen sind sehr anfällig auf Spinnmilben. Auf der Blattunterseite müssten winzige Tierchen erkennbar sein und feine Spinnweben in den Blattachsen. Die Blätter wirken durch das Saugen des Schädlings farblos und matt. Holen Sie in einem Fachhandel ein Mittel gegen Spinnmilben. Behandeln Sie die Pflanze mit Wiederholungen. Wichtig: vor der Behandlung so viele Blätter wie möglich abschneiden.*

### B.4. TRANSLATION OF EDITED INPUT

*Hello plant doctor. I bought a banana plant and a fan palm a few months ago. Together with 2 additional palm trees (the/that one is a dracaena marginata.) The leaves of the banana plant and the fan palm become, however, more dry and drier now for some time and die. However, the plants are sufficiently moist. Is that due to the humidity of the air? If so: How can I increase these and counteract this? Thank you!*
*The plant doctor:*
*A lack of humidity of the air can be the reason for the brown leaves and edges. If the leaves dry up, that has other reasons probably. A possibility would be spider mites. Bananas are very susceptible to spider mites. On the underside of the leaf(s), minute small animals had to be recognizable and fine cobwebs in the leaf axles. The leaves appear colorless and dull through the sucking of the pest. Buy a remedy against spider mites at a specialized dealer. Treat the plant with repetitions. Important: Cutting off as many leaves before the treatment as possible.*

## C. English to French Translation

### C.1. ORIGINAL INPUT

*Fighting Houseplant Pests*
*Everyone who has indoor plants must, at one time or another, deal with pests. If you notice that one of your plants is infested, isolate it immediately and identify the pest. If you're not sure what it is pinch off a leaf and bring it into XYZ. We can more than likely identify the little critter for you. Please place the leaf or leaves in a plastic bag, we don't want your pest getting out and running all over XYZ! Inspect your plants often to make sure those little darlings don't take a good hold before you can get them eradicated!* (. . . ) *A magnifying glass comes in real handy when checking for pests.*

### C.2. TRANSLATION OF ORIGINAL TEXT

*Parasites de Combat de Houseplant*
*Chacun qui a les plantes d'intérieur doit, en même temps ou des autres, affaire avec des parasites. Si vous notez qu'une de vos plantes est infestée, isolez-l'immédiatement et identifiez le parasite. Si vous n'êtes pas sûr ce que c'est pincement outré d'une feuille et ne l'introduisez pas dans XYZ. Nous mettons en boîte identifions plus que probablement la petite créature pour vous. Veuillez placer la feuille ou des feuilles dans un sachet en plastique, nous ne voulons pas votre parasite sortant et exécutant partout XYZ ! Examinez vos plantes souvent pour s'assurer que ces petits chéris ne prennent pas une bonne prise avant que vous puissiez les obtenir supprimées !* (. . . ) *Une loupe vient dans vrai maniable en vérifiant les parasites.*

### C.3. EDITED INPUT

*How to Fight Houseplant Pests*
*Everyone who has indoor plants must, at one time or another, deal with pests. If you notice that one of your plants is infested, isolate it immediately and identify the pest. If you're not sure what it is, pinch off a leaf and bring it into XYZ. We can more than likely identify the little critter for you. Please place the leaf or leaves in a plastic bag. We don't want your pest getting out and running all over XYZ! Inspect your plants often in order to make sure that those little darlings don't take a good hold before you can get them eradicated!* (. . . ) *A magnifying glass is very useful when you check for pests.*

### C.4. TRANSLATION OF EDITED INPUT

*Comment combattre des parasites de Houseplant*
*Chacun qui a les plantes d'intérieur doit, en même temps ou des autres, affaire avec des parasites. Si vous notez qu'une de vos plantes est infestée, isolez-l'immédiatement et identifiez le parasite. Si vous n'êtes pas sûr ce qui est elle, pincez outré d'une feuille et introduisez-la dans XYZ. Nous mettons en boîte identifions plus que probablement la petite créature pour vous. Placez s'il vous plaît la feuille ou des feuilles dans un sachet en plastique. Nous ne voulons pas votre parasite sortant et exécutant partout XYZ ! Examinez vos plantes souvent afin de s'assurer que ces petits chéris ne prennent pas une bonne prise avant que vous puissiez les obtenir supprimées !* (. . . ) *Une loupe est trés utile quand vous vérifiez les parasites.*

## Notes

[1] This paper is a revision of the material that the authors covered in their tutorial on MTranslatability at AMTA-2000 in Cuernavaca, Mexico.

[2] This section draws on many sources as well as our own research. While we cannot acknowledge each contribution individually, we would like to mention the following papers as major sources: Kohl (1999), Language Partners International (2000), Korpela (1998), Harkus (2000), Mitamura (1999).

[3] All the translations in this paper, with the exception of the translations provided in (17b), (19), (31), and (32), are produced by one of the four MT systems that we tried out. We have not identified the MT systems because an evaluation is not the purpose of this paper. Suffice it to say that some MT systems are significantly more robust with respect to some of these issues than others and that, obviously, some translations are significantly better than others. For the purposes of this paper, we chose translations not for the quality of their output, but for the difference in translation before and after a rewrite. Hence the translations may still not be perfect after a rewrite, even though they did benefit from the rewritings.

[4] In the gloss for Spanish, "imp-refl" is used to indicate the special use of the reflexive *se* in impersonal constructions.

[5] Winograd's original example (i)–(ii), as stated in Winograd (1972), does not illustrate as clearly the need for pronoun resolution:

(i) The city councilmen refused the demonstrators a permit because they feared violence.

(ii) The city councilmen refused the demonstrators a permit because they advocated revolution.

[6] Notice the similarity to example (16b).

[7] Zar's poem is widely reproduced on the World-Wide Web, without acknowledging his authorship, and with the alternative title "An owed to the spell chequer".

[8] This is the term Haller uses; his example suggests that he means "*ing*-words" in our sense.

[9] Underwood and Jongejan (2001) report on a tool to assess the MTranslatability of sentences and documents in the context of the European TQPro (Translation Quality for Professionals) project (Thurmair, 2000b). We will not attempt a description of their Translatability Checker here because apparently an evaluation of the effort is not yet available. Suffice it to say that the approach is similar to the tools described here in that it gathers so-called translatability indicators and deducts points from a perfect score accordingly.

[10] Hutchins and Somers (1992: 94) make the very good point that it does not really matter what kind of ambiguity the system is up against; what matters is whether the system has the relevant data for disambiguation. And this we cannot be sure of when composing a text independently of a specific system.

[11] Hutchins (1998) reports on a "description by Peter Pym in 1988, of the successful use of the Weidner system at Perkins Engines. Some time before, the company had sought to improve the quality of its technical documentation by introducing a "simplified" language for authors, PACE (Perkins Approved Clear English). The use of a "controlled" language with Weidner proved a great success – Pym could report major savings in translation costs, particularly when texts were to be translated into a number of languages – and demonstrated to many sceptics that MT was a realistic option even for relatively small organizations."

[12] The difference between 0 and 1 is that 1 indicates a completely garbled translation, whereas 0 indicates an intelligible translation that does not reflect the meaning of the source.

[13] This is also shown in the ratings of the German translation of the sentence (iii)–(iv)

(iii) When and how should I prune them?

to

(iv) Wann und wie sollte ich sie beschneiden?

While two evaluators rated the translation as 4 and 5 (from a 3 without editing), the third evaluator rated the original and the edited sentence both as 0. In his idiolect, the German verb *beschneiden* means primarily 'circumcise', whereas in the other evaluators' use of the language, it means both 'prune' and 'circumcise' (among other possible meanings).

## References

AECMA: 1995, 'A Guide for the Preparation of Aircraft Maintenance Documentation in the Aerospace Maintenance Language. AECMA Simplified English'. Brussels. AECMA Document: PSC-85-16598, Issue 1.

Anonymous: 1995, 'Why Spell Check Does not Work – A Linguistic Odyssey'. Available on the Web at http://www.linguistlist.org/issues/6/6-407.html (as of May 9, 2001). Linguist List 6.407.

APS: 1997, 'Dear Plant Doctor...'. Available on the Web at http://www.scisoc.org/visitors/plantdoc2.htm (as of April 4, 2002). American Phytopathological Society.

Becker, M., A. Bredenkamp, B. Crysmann, and J. Klein: forthcoming, 'Annotations of Error Types for German USENET News Corpus'. In: A. Abeillé (ed.): *Treebanks. Building and Using Syntactically Annotated Corpora*. Dordrecht, Kluwer. To appear.

Bernth, A.: 1997, 'EasyEnglish: A Tool for Improving Document Quality'. In: *Fifth Conference on Applied Natural Language Processing*. Washington, DC, pp. 159–165.

Bernth, A.: 1998, 'EasyEnglish: Preprocessing for MT'. In: *Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA, pp. 30–41.

Bernth, A.: 1999a, 'A Confidence Index for Machine Translation'. In: *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*. Chester, England, pp. 120–127.

Bernth, A.: 1999b, 'Controlling Input and Output of MT for Greater User Acceptance'. In: *Translating and the Computer 21*. London, no page numbering.

Bernth, A.: 1999c, 'Tools for Improving E-G MT Quality'. In: *Workshop on Problems and Potential of English-to-German MT Systems*. Chester, England, no page numbering. Held in conjunction with the 8th International Conference on Theoretical and Methodological Issues in Machine Translation.

Bernth, A. and C. Gdaniec: 2000, 'A Translation Confidence Index for English-German MT'. Technical Report RC 22403, IBM T.J. Watson Research Center, Yorktown Heights, NY.

Bernth, A. and M. C. McCord: 2000, 'The Effect of Source Analysis on Translation Confidence'. In: J. S. White (ed.): *Envisioning Machine Translation in the Information Future, 4th Conference of the Association for Machine Translation in the Americas*. Berlin, Springer, pp. 89–99.

Bredenkamp, A., B. Crysmann, and M. Petrea: 2000, 'Looking for Errors: A Declarative Formalism for Resource-Adaptive Language Checking'. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation*. Athens, Greece, pp. 667–673.

DFKI: 2000, 'FLAG: Flexible Language and Grammar Checking', Available on the Web at http://flag.dfki.de (as of May 17, 2001).

Ducrot, D. M.: 1989, 'Le système TITUS IV: système de traduction automatique et simultanée en quatre langues' [The Titus IV system: a system for simultaneous machine translation in four languages]. In: A. Abbou (ed.): *Traduction assistée par ordinateur: perspectives technologiques, industrielles et économiques envisageables à l'horizon 1990: l'offre, la demande, les marchés et les évolution en cours*. Paris, Editions Daicadif.

Elliston, J. S. G.: 1979, 'Computer Aided Translation: A Business Viewpoint'. In: B. M. Snell (ed.): *Translating and the Computer*. Amsterdam: North-Holland, pp. 149–158.

Farwell, D., L. Gerber, and E. Hovy (eds): 1998, *Machine Translation and the Information Soup, Third Conference of the Association for Machine Translation in the Americas*. Berlin, Springer.

Gdaniec, C.: 1994, 'The Logos Translatability Index'. In: *Technology Paratnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in The Americas*. Columbia, Maryland, pp. 97–105.

Godden, K.: 1998a, 'Controlling the Business Environment for Controlled Language'. In: *Proceedings of the Second International Workshop on Controlled Language Applications*, CLAW 98. Pittsburgh, PA, pp. 185–190.

Godden, K.: 1998b, 'Machine Translation in Context'. In: Farwell et al. (1998), pp. 158–163.

Godden, K.: 2000, 'The Evolution of CASL Controlled Authoring at General Motors'. In: *Proceedings of the Third International Workshop on Controlled Language Applications*, CLAW 2000. Seattle, WA, pp. 14–19.

Godden, K.: 2002. Personal communication.

Haller, J.: 2000, 'MULTIDOC. Authoring Aids for Multilingual Technical Documentation'. In: *Proceedings of the 1st Congress of Specialized Translation*. Barcelona, pp. 143–147.

Halliday, M. A. K. and R. Hasan: 1976, *Cohesion in English*. London, Longman Group.

Harkus, S.: 2000, 'Writing for Translation'. In: *Proceedings of the Australasian Online Documentation Conference*, Brisbane, Australia, pp. 154–166.

Hashida, K.: 1997, 'Global Document Annotation'. In: *4th Natural Language Processing Pacific Rim Symposium '97*, NLPRS-97. Phuket, Thailand.

Hashida, K.: 2000, 'GDA: Semantically Annotated Documents as Intelligent Content'. In: *Coling 2000 Workshop on Semantic Annotation and Intelligent Content*. Luxembourg. Oral presentation only.

Hoard, J., R. Wojcik, and K. Holzhauser: 1992, 'An Automated Grammar and Style Checker for Writers of Simplified English'. In: P. Holt and N. Williams (eds): *Computers and Writing. State of the Art*. Oxford, Intellect, pp. 278–296.

Holmback, H., L. Duncan, and P. Harrison: 2000, 'A Word Sense Checking Application for Simplified English'. In: *Proceedings of the Third International Workshop on Controlled Language Applications, CLAW 2000*. Seattle, WA, pp. 120–133.

Huijsen, A. S.-O.: 1998, 'Controlled Language – An Introduction'. In: *Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA, pp. 1–15.

Hutchins, W. J.: 1998, 'Twenty years of Translating and the Computer'. In: *Translating and the Computer 20*. London, no page numbering.

Hutchins, W. J. and H. Somers: 1992, *An Introduction to Machine Translation*. London, Academic Press.

Hydroplant AG: 2001, 'Frage von Jonas 2001.12.20 [Question from Jonas]'. Available on the Web at http://www.hydroplant.ch/pflanzendoktor/ (as of April 4, 2002).

Jackendoff, R. S.: 1972, *Semantic Interpretation in Generative Grammar*. Cambridge, Massachusetts, MIT Press.

Kamprath, C., E. Adolphson, T. Mitamura, and E. Nyberg: 1998, 'Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English'. In: *Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA, pp. 51–61.

Kay, M.: 1973, 'The MIND System'. In: R. Rustin (ed.): *Natural Language Processing*. New York, Algorithmics Press, pp. 155–188.

Kohl, J. R.: 1999, 'Improving Translatability and Readability with Syntactic Cues'. *TechnicalCOMMUNICATION*, pp. 149–166.

Korpela, J.: 1998, 'Translation-friendly authoring, especially in HTML for the WWW'. Available on the Web at http://www.malibutelecom.com/yucca/transl/ and http://www.cs.tut.fi/jkorpela/transl/ (as of May 17, 2001).

Kurohashi, S. and M. Nagao: 1994, 'A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures'. *Computational Linguistics* **20**, 507–534.

Kurohashi, S. and M. Nagao: 1998, 'Building a Japanese Parsed Corpus while Improving the Parsing System'. In: *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain, pp. 719–724.

Langlais, P., G. Foster, and G. Lapalme: 2000, 'Unit Completion for a Computer-aided Translation Typing System'. *Machine Translation* **15**, 267–294.

Langlais, P., G. Lapalme, and M. Loranger: 2002, 'TransType: from an Idea to a System'. *Machine Translation* (Special Issue on Embedded Machine Translation Systems). To appear.

Language Partners International: 2001, 'Writing for Translation'. Available on the Web at http://www.languagepartners.com/reference-center/wri4tr.htm (as of May 17, 2001).

Maruyama, H., H. Watanabe, and S. Ogino: 1990, 'An Interactive Japanese Parser for Machine Translation'. In: *COLING-90: Papers presented to the 13th International Conference on Computational Linguistics*, Helsinki, Vol. 2, pp. 257–262.

Mason, J. and R. Rinsche: 1995, *Translation Technology Products*. London, OVUM Ltd.

McCord, M. C.: 1980, 'Slot Grammars'. *Computational Linguistics* **6**, 31–43.

McCord, M. C.: 1990, 'Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars'. In: R. Studer (ed.): *Natural Language and Logic: International Scientific Symposium*. Berlin: Springer Verlag, pp. 118–145.

McCord, M. C. and A. Bernth: 1998, 'The LMT Transformational System'. In: Farwell et al. (1998), pp. 344–355.

Means, L. and K. Godden: 1996, 'The Controlled Automotive Service Language (CASL) Project'. In: *Proceedings of the First International Workshop on Controlled Language Applications, CLAW-96*. Belgium, Leuven, pp. 106–114.

Means, L. G., P. Chapman, and A. Liu: 2000, 'Training for Controlled Language Processes'. In: *Proceedings of the Third International Workshop on Controlled Language Applications, CLAW 2000*. Seattle, WA, pp. 1–13.

Mitamura, T.: 1999, 'Controlled Language for Multilingual Machine Translation'. In: *Machine Translation Summit VII: MT in the Great Translation Era*. Singapore, pp. 46–52.

Mitamura, T. and E. Nyberg: 1995, 'Controlled English for Knowledge-Based MT: Experience with the KANT System'. In: *Proceedings of Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI 95*. Leuven, Belgium, pp. 158–172.

Mitamura, T. and E. Nyberg: 2000, 'Controlled Languages'. Technical report, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA. AMTA 2000 Tutorial.

Multilint: n.d. Available on the Web at http://www.iai.uni-sb.de/en/multien.html (as of May 17, 2001).

Nyberg, E. H. and T. Mitamura: 1996, 'Controlled Language and Knowledge-Based Machine Translation: Principles and Practice'. In: *Proceedings of the First International Workshop on Controlled Language Applications, CLAW 96*. Leuven, Belgium, pp. 137–142.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik: 1972, *A Grammar of Contemporary English*. London, Longman.

Reuther, U.: 1998, 'Controlling Language in an Industrial Application'. In: *Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA, pp. 174–184.

Reuther, U. and A. Schmidt-Wigger: 2000, 'Designing a Multi-Purpose CL Application'. In: *Proceedings of the Third International Workshop on Controlled Language Applications, CLAW 2000*. Seattle, WA, pp. 72–82.

Schmidt-Wigger, A.: 1998, 'Grammar and Style Checking for German'. In: *Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA, USA, pp. 76–85.

Schütz, J.: 2001, 'Ontologies in Terminology Work. Enabling Controlled Authoring'. In: F. Steurs (ed.): *Terminology in Advanced Microcomputer Applications; Sharing Terminological Knowledge; Terminology for Multilingual Content; (TAMA-2001)*. TermNet Publisher, Vienna, no page numbering.

Shehadeh, C. M. E. and J. B. Strother: 1994, 'The Use of Computerized Readability Formulas: Bane or Blessing?'. In: *Proceedings of the Society for Technical Communication Annual Conference*. Minneapolis, MN, pp. 225–227.

Smart Communication Inc.: 1998, 'MAXit: The SMART Expert Editor'. In: *Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA, p. 196.

Thurmair, G.: 1997, 'Exchange Interfaces for Translation Tools'. In: *MT Summit VI: Machine Translation Past Present Future*. San Diego, CA, pp. 74–94.

Thurmair, G.: 2000a, 'Text Handling Standard: OTEXT V323'. Translation Quality for Professionals (TQPro). Available on the Web at http://www.tqpro.de/dorsexternal/Otext-V323.pdf (as of May 18, 2001).

Thurmair, G.: 2000b, 'TQPro: Quality Tools for the Translation Process'. In: *Translating and the Computer 22*. London, no page numbering.

Tomita, M.: 1986, 'Sentence Disambiguation by Asking'. *Computers and Translation* **1**, 39–52.

Underwood, N. L. and B. Jongejan: 2001, 'Translatability Checker: A Tool to Help Decide Whether to Use MT'. In: *MT Summit VIII: Machine Translation in the Information Age*. Santiago, de Compostela, Spain, pp. 363–368.

Watanabe, H., K. Nagao, M. McCord, and A. Bernth: 2000, 'Improving Natural Language Processing by Linguistic Document Annotation'. In: *Coling 2000 Workshop on Semantic Annotation and Intelligent Content*. Luxembourg, pp. 20–27.

Whitelock, P., M. M. Wood, B. Chandler, N. Holden, and H. Horsfall: 1986, 'Strategies for Interactive Machine Translation: the experience and implications of the UMIST Japanese project'. In: *11th International Conference on Computational Linguistics, Proceedings of Coling 86*. Bonn, pp. 329–334.

Winograd, T.: 1972, *Understanding Natural Language*. New York, NY, Academic Press.

Wojcik, R. H. and J. Hoard: 1997, 'Controlled Languages in Industry'. In: R. A. Cole (ed.): *Survey of the State of the Art in Human Language Technology*. Cambridge, Cambridge University Press, pp. 238–239.

Wojcik, R. H., J. Hoard, and K. Holzhauser: 1990, 'The Boeing Simplified English Checker'. In: *Proceedings of International Conference. Human Machine Interaction and Artificial Intelligence in Aeronautics and Space*. Toulouse, France, pp. 43–57.

Wojcik, R. H. and H. Holmback: 1996, 'Getting a Controlled Language Off the Ground at Boeing'. In: *Proceedings of the First International Workshop on Controlled Language Applications, CLAW 96*. Leuven, Belgium, pp. 22–31.

Wojcik, R. H., H. Holmback, and J. Hoard: 1998, 'Boeing Technical English: An Extension of AECMA beyond the Aircraft Maintenance Domain'. In: *Proceedings of the Second International Workshop on Controlled Language Applications, CLAW 98*. Pittsburgh, PA, pp. 114–113.

Zar, Jerrold H.: 1994, 'Candidate for a Pullet Surprise', *Journal of Irreproducible Results*, Jan/Feb, p. 13; see also http://tenderbytes.net/rhymeworld/feeder/teacher/pullet.htm (as of May 13, 2002).