Machine-friendly or reader-friendly?

Two terms with which I became acquainted when using machine translation in Germany to translate hardware and software manuals were 'reader-friendly' (leserfreundlich) and 'machine-friendly' (maschinenfreundlich). But while the scope of reader-friendly and machine-friendly can be co-extensive, not everything which is reader-friendly is machine-friendly. Even more the case: not everything which is machine-friendly is reader-friendly. The notion that the terms are interchangeable has been encouraged by proponents of MT wishing to dismiss genuine weaknesses in automatic language analysis as errors prompted by weak input and therefore not the fault of MT.

Dr William J. Niven

What I will do in this article is to examine, by example of the German language, the degree to which reader-friendly and machine-friendly converge and diverge: something will be revealed of the seemingly insuperable deficiencies of MT.

Format

These deficiencies become apparent already at the earliest stage in the MT process, namely deformatting. Text in technical manuals never comes as a seamless stream of sentences. Rather it is presented in the form of sections, subsections, paragraphs, columns, tables and lists. Clearly this format information is not required in the actual analysis and translation of the text, and so has to be stripped away. But it should not be simply discarded, and systems such as METAL are able to store format information and then reimpose it upon the translated text. Unfortunately, however, the process of text extraction can be error-prone, especially when dealing with German technical manuals, which are usually heavily formatted. The example below represents a typical distribution of information in a German technical manual:

Sie müssen immer darauf achten, daß nach Abschluß Ihrer Arbeit

- der Bildschirm UND
- das Terminal

ausgeschaltet werden.

An MT system will in all probability recognise four separate sentences here, not one. A series of blanks is usually interpreted as an end-of-sentence-marker. As a result there is no chance of the four lines of text being considered as one syntactical unit, and the resulting analysis and translation, based as they are on chunks, are bound to be very poor. A second problem in the above example is the use of uppercase ("UND"). Uppercase words are generally treated as unknown nouns (unless coded), working on the principle that German is lowercase except where acronyms or product names are meant (e.g. "Der Computerriese ABM"). The UND here would therefore be perceived as nominal and the conjunctional use completely missed.

There is a strong case to be made for the reader-friendliness of this format. The highly formatted presentation of material always induces the reader to read more slowly, and the physical separation of certain words from the rest of the sentence by a blank line serves to highlight and draw the reader's attention to them.

Formatting can also mean using truncated sentences, with, say, one subordinate clause serving a string of main clauses. This is particularly common in tables:

Wenn das System nach dem Einschalten nicht hochläuft, dann könnte es daran liegen, daß	1) der Stecker am Prozessor locker sitzt; 2) der Prozessor defekt ist; 3) der Einschaltmechanismus nicht funktioniert

An MT system, in so far as it interprets each column of a table as an independent structure and reads vertically, has no chance of recognising that the right-hand column contains a sequence of continuations of the sentences started in the left-hand column. It will divide up the linguistic information within these two columns into four translation units, none of which, when analysed, can possibly render a complete sentence: the first translation unit will be composed of an antecedent subordinate clause with incomplete main clause (", daß..."), while the following three units will consist of clauses with final-position verb but no preceding subordinate clause and linking conjunction. The resulting translation will be very poor. Nor would it help much if the MT system were to read horizontally across column boundaries: line one of the right-hand column is a continuation of line three of the left-hand column, not of line one, which would be the assumption of a horizontal reading. The only way to produce a clearly machine-friendly version of the above would be to dispense with the table and rewrite in complete sentences.

Again, this may be reader-friendly as well, but the tabular form arguably provides a more logical structuring and clearer distribution of the information. As a general rule, machine-friendly, given the problems a machine has correctly understanding the relationship between format and text, strives for an essentially format-free input in which information is arranged in the form of continuous text or "Fließtext"; by contrast, reader-friendly is coming more and more to mean the formally sophisticated organisation of linguistic material within tables, graphs, insets etc., exploiting all the modern resources of automatic text processing such as tabs, indentation, fluctuating margins and, last but not least, blanks and return characters.

Analysis

Once an MT system has extracted the linguistic information from the formal layout and formed sentence units, it is faced with the problem of analysing and translating these. Only where the units consist of correct sentences is there a possibility of accurate analysis. MT systems are absolutely dependent upon flawless input. Printing errors and spelling mistakes can usually be identified and then pre-edited out by using conventional spelling-checkers or relying on preanalysis (during preanalysis the program compares the strings in the text with the words in the computer's source-language lexicon and identifies all unknown sequences). But MT systems will not pick up other typical source input errors: words accidentally omitted from a sentence, for instance, or words accidentally repeated, cannot normally be traced prior to analysis, where they cause havoc. Syntactical and grammatical errors also make accurate analysis impossible. Successful automatic language analysis requires the correct application of those language generation rules with which it has been provided. A human being is at an advantage here, being able in many cases to reconstruct the *intended* sentence, mentally adjusting faulty grammar or supplying missing commas. However, it can hardly be contested that correct input would be as much of a blessing for a human reader as for an MT system. The human capacity for recognising discrepancies between the actual and intended in sentences is not a universal remedy: some input errors are intractably opaque or ambivalent, quickly exhausting the patience of the human reader.

Both MT and human reader will also clearly benefit from the use of short or at least medium length sentences (as a rule of thumb, not more than 20 words). The more information contained in a sentence, the more the MT system or reader has to take in: considerable demands are made on memory. And there is always a risk of the relationship between the parts of the sentence becoming obscure the longer the sentence continues. Long sentences, of course, are often caused by the use of colons and semi-colons. Such sentences are not necessarily difficult for a reader, nor *need* they be for MT. Colons are usually interpreted by MT systems as indicating boundaries between syntactically and semantically self-contained units, a fairly safe assumption except when colons are used within parenthetical constructions1. The same assumption is made in the case of semi-colons. Here, while still being largely tenable, the assumption seems less safe: semi-colons can be used to divide main and subordinate clause. The biggest difficulty is posed rather by long sentences composed of intertwined strings of subordinate and relative clauses separated by a superabundance of commas. The following sentence provides a good illustration of this:

Der Symbolabstand, der den Abstand, der sich aus der Symboldefinition ergibt, erhöht, kann durch Eingabe eines Kommandos in der DOS-Shell verändert werden.

Nested within the main clause is a relative clause, within which another relative clause is nested: the higher the nesting rate, the greater the likelihood of a mechanical analysis failing to disentangle the intertwined clauses correctly. The decision to nest clauses rather than append one to the other is essentially a stylistic decision and thus a rhetorical, never a logical one: logical syntax annexes clauses, assuming that relations between sense units are sequential or at least additive. Given that technical writers are describing processes, logical syntax is a *sine qua non* of technical manuals. Such sentences not only cause MT systems problems, they also force the human reader to read sentences several times and mentally re-order the cluttered clauses: though a human being can generally, after some effort, perform a successful analysis, whereas a machine will frequently fail.

Of course one can nest not only clauses, but also information within clauses. Left-branching prenominal adjective constructions, often using participials, are ideal for this in German:

Eine auf allen Systemen ablauffähige und auf die jeweiligen Anforderungen zugeschnittene Anwendung.

Here, the article and noun are separated by ponderous prepositional phrases, both nominal and adjectival. An MT system will have problems recognising the syntactically attenuated link between the determiner "eine" and "Anwendung". Without doubt the phrase would also cause a human reader similar difficulties to those entailed in reading intertwined clauses. Simplification of the syntax and division into several sentences would be of all-round benefit:

Die Anwendungen dieses Moduls laufen auf allen Systemen. Sie können auch auf die jeweiligen Anforderungen zugeschnitten werden.

There are many causes for long sentences. In literary texts, they are perfectly permissible: there the author is answerable only to himself or herself. But the writer of technical documentation is answerable only to the end-user and is thus duty bound to package information in such a way that it is easily and swiftly understood. Brevity is essential: information must be passed on in manageable portions, frequent full-stops allowing the reader those mental and visual pauses so vital in the process of absorbing new information.

Unfortunately, however, writers of technical documentation do not always do what is expected of them. That writers of technical documentation are frequently unable to present material clearly can be explained in one of two ways. Either they have literary ambitions and are more interested in drawing elegant literary arabesques around the subject-matter than in stripping it to its bare essentials — or they are themselves unsure how the product they are supposed to be describing works. Rarely do technical developers write their own manuals. Most have neither the time nor inclination, and often they do not trust their own ability to describe their own inventions in finished form. So they pass on often complex notes together with technical specifications to the writers of documentation, most of whom have arts degrees and little natural aptitude for grasping technical concepts. This potent mixture of inadequate information and lack of aptitude can be overcome by frequent phonecalls between technical developers and writers, but often it is not overcome, and the result is a poor manual.

In my experience many long sentences result not so much from the stylistic deficiencies of the technical writers as from the fact that they lack product knowledge. Short SVO sentences are not only syntactically simple, they are often *semantically* uncomplicated too: in other words, to write a short sentence you really have to know what you want to communicate. Those who do not know have recourse to longer sentences, in which adjectival, adverbial and prepositional phrases as well as subordinate and relative clauses can be used to distend and blur the information content. How an MT system or a normal human reader is supposed to extract sense from sentences which are not intended to yield it is of course the question here.

In this connection we have to consider the German passive, which flourishes in technical manuals and definitely is one of the causes of long sentences. Why? Nowadays the doing of the deed, the execution of a process, is more important than the issue of agency. We live in a world where technology semi-automatically or even automatically performs more and more of those things commonly done in the past by human labour.

The passive, one might argue, is modern precisely because it mirrors the notion of effortless automation. Because, especially in computer programs, certain features can be subjected to a whole chain of interlinked processing steps, passive clauses governed by one subject are concatenated to reproduce the notion of continuous processing and to avoid unnecessary repetition. Take the following example:

Wenn Sie die F-Taste drücken, wird die Datei unter dem von Ihnen vergebenen Namen gespeichert, in Ihr Verzeichnis zurückgeschrieben und dann alphabetisch unter alle anderen Dateien, die sich dort befinden, eingeordnet.

Noticeable here is that one auxiliary ("wird") governs three participles. A human reader will have no problem recognising the extended application of a single specification of subject and auxiliary, but MT systems often do. The main reason for this is that, if the sentence-level analysis fails (likely in the case of long sentences), a phrase-level analysis takes over, which means that only those phrases with explicit specification of subject and auxiliary will be half-satisfactorily analysed; those containing only a final-position participle will be seen as syntactically incomplete and poorly translated. This problem in piecing together the parts of long passivised sentences is an example of the problem known in the machine-translation trade as "gapping".

Sentences containing passives do not need to be long to puzzle an MT system. Quite generally, translation programs often have difficulties distinguishing the agentive use of "von" and "durch" from other uses of these prepositions:

Die defekte Maschine wird von der Partnermaschine entkoppelt und zur Reparatur gebracht. Die Reparatur dauert zwei bis drei Tage.

Here, it is highly likely that the "von" introduces a prepositional phrase attached to "entkoppeln". The partner machine linked to the machine which has broken down is clearly not capable of taking it to be repaired, and is also unlikely to be doing the detaching. The probability in passive sentences of "von" or "durch" not being a passive marker is high: these prepositions ordinarily occur frequently in other functions. Such a likelihood is increased by the fact that there is no obligation to specify agency in passive sentences. Agency is often implied rather than specified. But while humans have a good chance of interpreting the significance of a preposition, MT is at a disadvantage.

The above example illustrated cases of passive sentences which might confuse an MT system but are unlikely to confuse a human being. But there are cases of passive structures which might also leave a human being feeling confused. This can happen quite frequently when a reader consults a manual to establish how something works and what his or her role is in making it work:

Schalten Sie den Computer ein. Jetzt wird das Programm geladen.

Here it is unclear whether the program is automatically loaded or whether the user has to load the program. One might think that the only way to prevent ambiguities of this kind arising in the first place is to specify agency using "von". Let us assume, for instance, that the program referred to is loaded automatically by a master program called ADL. All we need to do then is rewrite the above as follows:

Schalten Sie den Computer ein. Jetzt wird das Programm von ADL geladen.

But adding "von" together with a program name to the above sentence is counterproductive: the number of interpretative possibilities is thereby increased. Is the program being loaded by ADL? Or is a program called ADL being loaded? And if the latter is meant, who is doing the loading? Clear specification of agency by means of transforming passive into active structures, thereby forcing subject-object specification and elucidating grammatical relations, is advocated as the way out of this dilemma. The passage should read

Wenn Sie den Computer einschalten, lädt ADL das Programm.

As the Siemens-Nixdorf handbook *Empfehlungen für Fachtexte* puts it: "Aktive Formulierungen sind im Vergleich zu passiven konkreter, genauer und lebendiger: sie zwingen den Verfasser zu eindeutigen, klaren Aussagen".[2]

This brings us back to the issue of lack of clarity in technical documentation. While it is legitimate to use the passive to convey steps in what is clearly defined as an automated process, it is not legitimate to use it in such a way that it remains unclear whether a process *may* or *may not be* automatic. The reader is then left in the dark as to what the user does, and what is done by the program. Unfortunately this is precisely the aim of writers of technical documentation who are themselves unsure of agency in the processes they are describing. They take refuge in vague applications of the passive rather than risking active structures in which they would be compelled to clarify agency in a direction which might turn out to be erroneous. This is patently a misuse of the passive, a kind of semantic cowardice. The passive, as it were, reflects here a passivity of attitude, an irresponsible, if understandable indifference to truth (understandable because technical writers do work under extreme time pressure and are often unable to acquire further information from developers, who might be on holiday or difficult to reach for other reasons).

So some, but by no means all of the problems which MT systems have with the passive are shared by a human reader. The situation in the case of OVS word order is comparable. Putting the object before the verb is a syntactic possibility very typical of German, but much rarer in English.

Die Ameise frißt die Eidechse. Sie hat eine lange Zunge, womit sie die Ameise in den Mund nimmt.

Both nouns in the first sentence are preceded by an article which could be either the nominative or accusative form of the feminine definite article. Unfortunately, the nouns are spelt the same way regardless of whether they are accusative or nominative in case. Moreover, both nouns have the same semantic type: they are animate.

An MT system has thus no way of telling who is doing the eating, and will probably opt for a conventional SVO reading in its analysis. But we know that lizards generally eat ants, not vice versa; we possess empirical experience and world-knowledge inaccessible to MT programs. Verb framing which, in cases of ambiguity, allowed for the *larger* creature to be the subject (in which case nouns would need size markings) would have a chance here, but size is not always a reliable criterion for who eats whom. Flies, for instance, tend to prey on creatures much larger than themselves. No matter how finely-tuned the semantic categorisation of verbs and nouns is, it can never be refined enough to resolve all SVO/OVS ambiguities successfully. One could even imagine a context where the sentence "Die Ameise frißt die Eidechse" is preceded by a sentence stating that millions of soldier ants regularly attack and carry off geckos, in which case the ant may indeed be eating the lizard. Unlike human beings, who have the benefit of being able to read sentences as parts of a larger text, MT systems suffer in their analysis from the inability to apply rules across sentence boundaries. Vital extra-sentential contextual information is ignored. The second sentence of the above example, for instance, makes it quite clear who is the subject and who is the object in the first sentence. Such clues will not be picked up by an MT program.

Of course the contextual interpretative facility of the human reader is not unlimited: there are cases where OVS can result in ambivalences which remain unresolved for both human reader and MT systems. It is with such cases in mind that the Siemens-Nixdorf handbook recommends: "Um Mißverständnisse auszuschließen, sollte man diese Reihenfolge möglichst vermeiden".[3]

So far in this article I have considered problems posed by unwieldy sentences: these problems include cluttered clauses, phrase nesting, complicated passive and attributive constructions, and, finally, structural and semantic ambiguity. Primarily these are problems of analysis: MT systems, and human readers, must attribute verbs to their correct clauses, determiners to the correct noun, and decide in cases of ambivalence on the correct morphological or syntactic reading. Undoubtedly MT has more problems with attribution than a human being, though in cases of ambivalence both can be equally at a loss. If the analysis fails, of course, no good translation can result. But even where analysis is correct, translations can be poor. The mismatch of weak or even inaccurate translation to accurate analysis is a problem largely restricted to MT. The problem has a lexical and a structural or, more loosely speaking, *stylistic* dimension.

Translation

Sophisticated transfer lexicons in MT systems are capable of distinguishing between different meanings of the same word by means of context. Context can be syntactic: to take a simple example, the verb "erinnern" when used with a reflexive always means "to remember"; when used with non-reflexive direct objects, it means "to remind". Syntactic contextual information is, however, not always adequate: the verb "beziehen" when used with a direct object can mean several things, such as "to move into" (a flat) or "to subscribe to" (a newspaper). One can program an MT system so that it knows which meaning to choose when the canonical form of the object is "Zeitung" or "Wohnung", but because context is often supersentential and MT programs cannot make cross-sentential links between nouns and pronouns, the meaning of "beziehen" in the following passage will have to be guessed at:

Die Wohnung ist frisch gestrichen. Ich hoffe sehr, daß sie Ihnen gefällt. Sie kann sofort bezogen werden.

In the case of many German verbs and the majority of German nouns, there is often *no* syntactic context upon which to draw when seeking to distinguish meaning. It is at this point that MT systems have to rely on terminological subject-areas to help with semantic distinction. Thus the noun "Fehler" would be deemed to have different meanings according to the kind of manual in which it is used. In a manual describing hardware, it will mean "fault". In a manual describing software, it will mean "error". In a technical manual describing machines other than computers, it will be adequately translated as "defect". And in all other cases the best translation is probably "mistake". The tagging of meanings in the transfer lexicon as applicable to certain subject-areas — which have to be specified before carrying out the translation — is an excellent idea, but in practice, often proves sadly inadequate.

While technical manuals usually do deal with a particular subject-area, they are rarely so monolithic as never to make reference to other areas. Thus a manual describing the functionality of a particular processor or printer would fall into the subject-area category of hardware. Yet such a manual would certainly also describe, to a degree at least, the accompanying or appropriate software. Similarly, a manual describing a programming language is likely to contain descriptions of the hardware on which this language runs. The semantic classification for "Fehler" which I have just outlined does not take such interdisciplinary aspects into account. Depending on the selected tag, an MT system would only offer a consistent rendering as either "fault" (hardware tag) or "error" (software tag), which would soon prove too inflexible. And computer manuals of course refer not only to hardware and software problems, but to "Fehler" on the part of the user, in which case the translation "mistake" or "error" is to be preferred in some sentences. To argue that an MT system is consistent in its translations of terminology and therefore more reliable than a human translator — an argument frequently deployed by machine-translation advocates — misses the point that consistency is often undesirable.

Where MT systems do manage to provide a translation that accurately reproduces the sense of the sentence, it may still be a poor translation *stylistically* speaking. One of the complaints often made by post-editors of machine-translation output, most of whom are professional translators, is that machines provide over-literal translations which are wooden and stylistically unacceptable. What is meant here by over-literal is that MT programs by and large project the syntactical structures of the source language on to the target language. Given that German and English are often syntactically quite different in character, such a projection produces awkward and Germanicised English which the post-editor has to change. Translators with English as their mother tongue have even expressed the fear that constant exposure as post-editors to such Gerlish might in time so weaken their feeling for proper English prose that they will find themselves accepting much machine-translation output without qualms: a truly horrific thought.

The argument that over-literalism renders much machine-translation output unusable only holds if such output is intended for post-editing to a high standard. If it is intended merely to provide basic information about a particular product, then over-literalism is a difficulty that can be lived with. Certainly it *is* a difficulty.

As far as translating German into English is concerned, the problems posed by one-to-one syntactic mapping fall into four categories. The first problem is that German allows subordinate clauses to precede the main clause far more readily than the English language does. Take the following example:

Daß die Anzeige auf dem Bildschirm nicht an der richtigen Stelle erscheint, kann mehrere Gründe haben.
Literal translation: That the display on the screen does not appear in the right place can have several reasons.
Stylistically preferred translation: There can be several reasons why the display on the screen does not appear in the right place.

The second problem area is that German has a strong tendency towards nominalisations, especially deverbal nominalisations used in combination with semantically weak or empty verbs. In English these nominal structures would generally be rendered in verbal form. Thus:

Dann erfolgt die Löschung der Dateien.
Literal translation: Then the deletion of the files occurs.
Stylistically preferred translation: Then the files are deleted.

Nominalisations are often used with the passive:

Durch Angabe von REORG kann eine Neunumerierung vorgenommen werden.
Literal translation: By specification of REORG a renumbering can be undertaken.
Stylistically preferred translation: You can renumber by specifying REORG.

Here, the nominalisation plus verb has been replaced by one verb, and the passive has disappeared. Often, though not always, English prefers not just to denominalise in favour of verbs, but also to render German passives as active formulations. The fourth problem becomes apparent if we look again at the last example. In German, sentences frequently start with prepositional phrases (here the agentive "durch Angabe von..."); the introduction of the subject is delayed till later in the sentence. The stylistically preferred translation of this example moves the transmuted subject to the beginning of the sentence and shifts the prepositional phrase to the end. The following sentence provides another example:

Mit dieser Anwendung wird die Datei eröffnet.
Literal translation: With this statement the file is opened.
Stylistically preferred translation: The file is opened with (or: by entering) this statement.

The prepositional phrase, again agentive in scope, begins the sentence in the German, but English requires it to be post-positioned after the verb. The subject "file" begins the English sentence.

Clearly machine-friendly and reader-friendly here do not overlap, not even to an extent. All the German sentences in the last four examples are quite normal and valid syntactical compositions, easy to read and digest. While it would be machine-friendly to adjust the German syntax in order to render the literal English translation more fluid, adjustments would not be required as far as a reader is concerned. Such adjustments are, in any case, hardly a legitimate solution to the machine's problems. Approximating source text syntax to target text syntax might even, in the worst case, lead to a corruption of the source by the target syntax. The German would read like an overliteral translation of the English into which *it* is supposed to be translated.

The only hope of a valid solution lies with the programmers of MT. MT is capable, at word and phrase level, of extensive and potent lexical and syntactical transformations. What is needed is an expansion of this capability to allow for the syntactical transformation of whole sentences from the source to the target language. But even given the realisation of such a capability, any rules evolved are bound to be too rigid: not *every* passive in German becomes active in English, not *every* German subordinate clause must come after the main clause in English. Again, it is the feeling for style that often decides when to transform and when not. And MT systems will, alas, never develop an instinct for questions of style.

Conclusion

In conclusion it can be said that, as far as German-into-English translation is concerned, the applicability of the adjectives machine-friendly and reader-friendly is not co-extensive. Ultimately, MT systems have more problems analysing a language than a human reader. But there are areas where the reader would benefit from a reformulation as much as MT. Certainly it is to be hoped that the advent of MT and the resulting pressure brought to bear on technical writers to produce "translatable" documentation will yield results that are of benefit to man and machine.

Personally I believe that a greater transparency in the German style of many manuals is absolutely necessary. Only poetry, drama and novels should be characterised by a freedom in the handling of language and syntax. This freedom, in addition to criteria of imagination and originality, is in fact one of the defining elements of aesthetic writing. Technical writing on the other hand should ideally be much less ambitious in its use of language. The differences derive from the different purposes. The aim of creative writing is the expression of individuality. An aspect of this is the development of a style of language which is distinctly personal. Personal, here, means non-normed, going beyond and possibly even breaking with linguistic traditions. The aim of technical writing is the transmission of information. It is duty-bound to be non-metaphorical, object-related, precisely descriptive in a way poetic writing is not. Technical writing operates — or at least should operate — in accordance with a set of linguistic norms (brevity at phrase and sentence level, syntactical clarity, terminological consistency). The more technical writers actually adhere to these ideals, the more their manuals will be suitable for MT.

Notes

[1] In German manuals colons are often used as in the following example: *Diese Programme (Beispiel: OCIS) sind neu.* An MT system will separate this sentence at the colon into two independent units. Although it seems possible to solve this problem by providing the machine with the information that a colon after "Beispiel" within an opening parenthesis does not designate the end of a unit, or that an opening parenthesis cannot be separated from a closing parenthesis.

[2] *Verständlich und übersetzungsfreundlich schreiben: Empfehlungen für Fachtexte*, ed. Siemens-Nixdorf Language Services, Munich 1993, p.7.

[3] *Ibid.*, p. 13.