[*Translating and the Computer 24*: proceedings of the…International Conference…21-22 November 2002, London (Aslib, 2002)]

Translating tagged text - imperfect matches and a good finished job

Lorna Joy
Translation Team UK
SCHÜCO International KG
Whitehall Avenue, Kingston, Milton Keynes MK10 0AL
LJoy@schueco.com

Synopsis

This paper will highlight the issues relating to the translation of tagged text formats both from those considering taking on such work and for those commissioning it. In addition to its practical advice, it is also hoped to feed into the debate on how the value currently being placed on the work of translation professionals is diminishing due to the increasing (and sometimes overoptimistic) reliance on computer tools. It will indicate what is required when using translation memories with tagged text in its various forms. Using a case study of translating tagged text in producing technical documentation in an engineering company, it will offer advice on how to maximise matches, and offer some discussion on the issues surrounding the fair payment for tagged text translation.

Introduction

The advances in information technology have revolutionised both the speed and the variety of information presentation in a variety of media. The challenge of today is to satisfy the demand for that information in all its forms in a multitude of languages. The key to delivering on these demands is the translator. In this age of ever more sophisticated computer applications relating to the production of multilingual multimedia information flow, it is easy to forget that without the professional translator none of this would be possible.

In other words, such applications as exist, despite many claims to the contrary, are merely tools to help store, retrieve and expedite the production of printed or web-based information. However, it is surprising how often non-linguists are apt to view the tools as an end in themselves and expect the translator to fit the way they translate around the way the tools work, instead of using the tools to facilitate the optimum translation in a given context. Similarly, it is easy as a translator to be lulled into a false sense of what one tool or another can facilitate. And sadly, there is an increasing tendency to devalue the work of the translator in the mistaken belief that the tools are in some way doing the translator's work for them. It is this assumption I wish to challenge in this paper with particular reference to the translation of tagged text using translation memory.

The purpose of the presentation is to show how tags affect matches (or lack of them) from the TM and how to prevent the formatting exerting a detrimental effect on the finished translation. I shall refer mostly to the common tool (S-Tagger®) used in the translation of text from desktop publishing package (Interleaf/Quicksilver®), giving examples from their everyday use in a technical translation department. The approach will also be applied to the translation of other tagged formats with suggestions for producing a good finished job, regardless of medium.

Quality translation is produced by skilled translators

Though it may be regarded by some as stating the obvious, it is important not to forget the importance of the act of translating to a quality finished job. Using a compilation of other people's definitions I would like to begin by putting forward a definition which should take centre stage in any discussion of computer-assisted translation tools:

Translation is the transfer of meaning from one written language to another, so that it is entirely comprehensible in the context for which it was originally written, and reads as though it had been composed by a native speaking expert in the field.

The terms "meaning", "context" and "composed" are used intentionally to remind us that translating between two languages is one of the most complex higher order activities of the human brain. Therefore, concepts thrown up by computer tools of 90% accuracy or 70% matches, essentially numerical qualifiers, may actually be of limited **meaning** within the context of genuine translation. These arguments have frequently been used in discussion of machine translation, but in the face of an increasing perception that the use of translation memory tools will provide the golden button to produce translations automatically, it may legitimately also be applied here. It is not simply that the tools must be seen for what they are, an aid to translators in their work, but furthermore they also have their own problems. Nowhere is this more apparent than when the texts to be translated come complete with tags, in which case we are confronted with the two-fold problem of working out what to translate and how, and how best to store the segments in translation memory

Where does Tagged text come from and what is it for?

Tags are used, such as in HTML (Hypertext Markup Language) to give layout and formatting information which convert into the layout on a web-page. Similarly computer programmes such as the Trados® S-Tagger were derived to extract the text from documentation prepared using desktop publishing (DTP) packages such as Framemaker® and Interleaf®, in order that the text may be processed away from other distractions such as pictures, graphics and non-linear layout. If, for example the same web page is required in another language, it makes sense to translate the text, and retain the same layout in both languages. Similarly, since documentation can now be produced as print-ready copy on a PC or workstation, it is possible (and may even be seen as essential from a corporate image point of view) to use the same layout complete with pictures and illustrations for multiple language versions. More recently, the development of XML for multi-purpose information flow, means that the same text may be converted using style sheets into a variety of formats, for use in print, in web presentation or even in dynamic web-based interaction.

Linear space saving versus "decontextualisation"

An engineering drawing, for example, might contain small units of text distributed randomly, at times even running vertically as well as horizontally, and any web page nowadays will consist of various units of text at different locations, and many additional related texts connected by hyperlinks. So, it would appear at first glance, that the linear form in which S-Tagged text and HTML are displayed would make life

easier for the translator. And indeed when compared with the laborious task of translating by overwriting texts in documents prepared in the DTP format, there are distinct advantages in being presented with the text in linear form.

At document level, however, the main problem of working with tagged text is what will be referred to here as "decontextualisation". The translator is presented with a text which may not even appear in the order it will appear on the printed page. The tags themselves may appear as apparently meaningless code in the middle of phrases (see below) but more significantly, each piece of text is a disembodied entity, without a context, so that great care is required to ensure that the resulting translation is not flawed. It has already been pointed out that context is an important factor in good translation, and if meaning is to be transferred at all, it must first be obvious from the source text what the meaning is. This may seem obvious, but it is surprising the number of people who forget that language is not "simply a nomenclature for a set of universal concepts"[1].

It should therefore be clear that for the translator to produce a quality document or web page in the target language, he or she must work with the final layout of the source as well as with the tagged text. It is also essential that the translator understands what the tags signify. Therefore, the minimum requirement when commissioning such texts is a copy of the pages or document in the source text as they are to be published, together with the tagged version for translation. In the case of HTML, access to the actual website complete with all its links (whether live on the web, or on a secure site), is essential to allow checking at every stage for contextual correctness.

The growing use of XML is likely to pose additional problems in this respect, as text segments may be used in any number of contexts, changing layout and using part texts in a variety of ways. Particularly when used in Content Management systems there will be a tendency to work in ever smaller translation units. Like building blocks, these units of text must be capable of being built back together in a variety of ways for a variety of purposes and still conveying the correct meaning in the correct context.

The danger in all tagged text translation is that there may be a tendency, owing to the disembodied nature of the text, to produce mechanistic translation, resulting in a literalness which impairs the transfer of meaning in a given context. This could become more pronounced if, due to reuse of previous texts, a translator is only to be given the bits that have been added when there is a new update of a company's content management repository. This then raises the question of how a translator is to be paid, and how much skill is required in producing a good end result. Is (s)he to be paid by the word for bite-sized chunks of text which someone later uses in a context yet to be decided? Or is (s)he to be viewed as a consultant, to be involved in ensuring quality translation for the organisation concerned? A closer look at translation unit level will highlight some of the problems which illustrate the unsuitability of the bite-sized chunk approach.

[1] Culler, J. (1976) *Saussure*, Glasgow Fontana/Collins

The two main problems with Tagged Text

There are two main areas for consideration when translating tagged text. One relates to the processing, the other to the storage of the tagged segments in the translation memory as meaningful and reusable translation units (TUs).

Editing tags and other pastimes!
Segmentation problems frequently arise in tagged text from DTP applications and are usually, although not always, a result of poor working practices in the use of the DTP. This type of problem will also be familiar to those who have used T-window to translate badly composed PowerPoint presentations. In the case study, this has proved a major problem, and highlights the fact that any tool is only as good as the person using it. A technical publisher whose aim is to produce an aesthetically pleasing effect in the original language may insert a hard return in otherwise justified text to prevent a word-break. Not only does this cut the sentence or phrase producing a meaningless segment, the hard return appears in the text as a tag <:hr>. Incorrectly used tabulations can produce the same effect (another problem experienced with amateur PowerPoint presentations) and will appear as the tag <:t>. At Schüco, this problem was overcome in-house by editing the original files before tagging.

The difference as far as the translator is concerned is that unlike the traditional approach when translating straight text, filtering text out from layout means that what the translator gets to translate looks something like this:

```
<ps "Seite" 1><:saf "Rahmen" 14>
<ps "Passer-Kreuz" 2><:iaf 15>
<ps "Kopf" 3><:t>TISCHBOHRMASCHINEN<:t><:t>Zubehör
Tischbohrmaschine BT<:sh>15<:t>
<ps "S_Rahmen" 22><:af 16>
<af 16>
<tr 31>
<ps "text Masch." 23>Maschinenständer
<ps "text Masch." 24>für Tischbohrmaschine BT<:sh>15 (Art.<:sh>Nr.
293 965).
<ps "text Masch." 25><:fc 5>D<:t><:/fc>Ablage für Werkzeuge
</tr>
<tr 42>
<ps "Absatz" 19>Technische Daten:
```

Fig 1. S-Tagged text from Quicksilver/Interleaf ® - how it appears using Word

It is also possible to process the text using the Trados Tag Editor (similar tools are also available with other Translation Memory packages). The tags are then more neatly configured as shown in Fig. 2, but they still need to be understood, (e.g. <:fc 5> means some kind of font change), and manipulated. It is important to know that every font change requires a closer <:/fc>, as one without the other will result in a error, and prevent reconversion to the original format. The tags can be locked, so that the

translator needn't touch them, but in practice they **have** to be edited. Not only must the translator be aware of what these tags mean, they must be able to remove them, and even insert them (e.g. when a word is hyphenated in English where it is not in German). German inverted commas or the Spanish inverted question mark may appear as a tag, which must be removed in other language versions. Similarly, a font change will appear if a section of text is highlighted for emphasis, but to produce the same emphasis may require the emboldening of more than one word, and will almost certainly occupy a different position in the sentence in the target language.

So the idea that the translator can simply translate the words between the tags is not quite the whole story. Unless the translator knows what the tags signify, (s)he cannot use them, resulting in the need for more extensive reformatting and the possibility of incorrect emphasis being placed on certain text, incorrect punctuation, and needless time being spent on reconversion of Tagged texts which contain multiple errors.

In a case where extensively formatted text was tagged and sent to an external translator (with copies of the original document as reference and having given a demonstration of working with tagged text), not only was the text returned with tags damaged and missing, the translator was unable to work out what to do when the Tag Verifier (a tool which checks the tags in the file being edited with a copy of the original) showed up the errors. The whole episode involved the in-house staff in additional work. The large amounts of time spent by translators manipulating tags when they could be translating is not helped by the assumption on the part of those commissioning the translation that such work deserves a lesser payment!

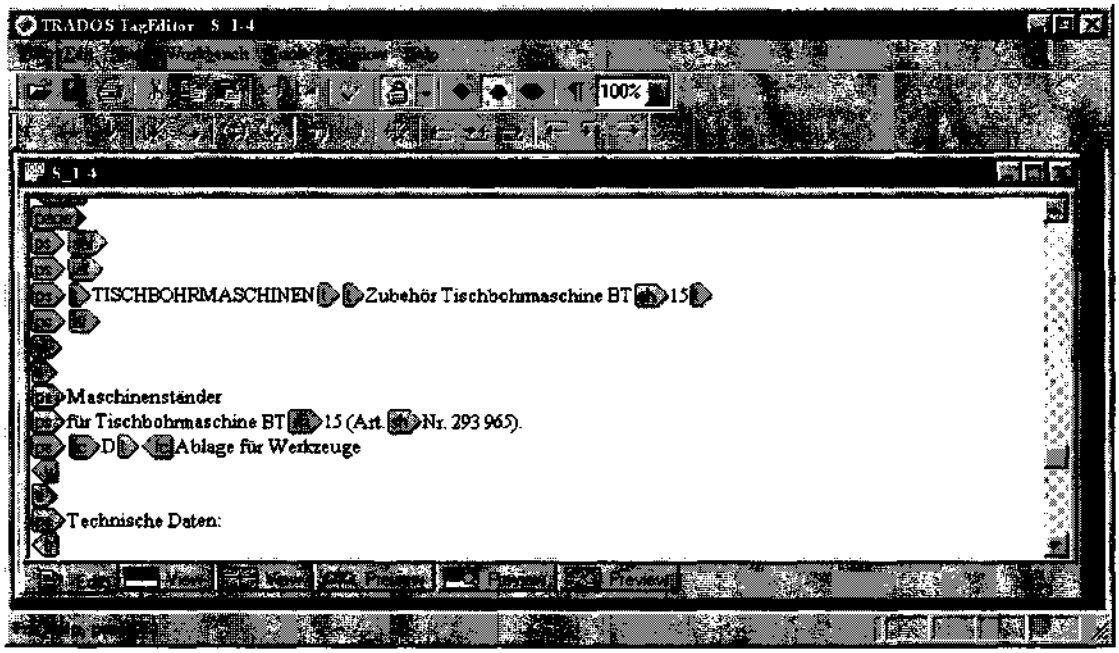


Fig. 2 S-Tagged text from an Interleaf® document viewed in Tag Editor

Storage and Retrieval - The Trouble with Tags!

Even in the most perfectly formulated document, translation of the tagged text using translation memory is still fraught with difficulties. In particular, the difficulty of obtaining matches; the prevalence of apparent matches which are in fact mismatches; and last but not least, the propensity Workbench® seems to have in misplacing the

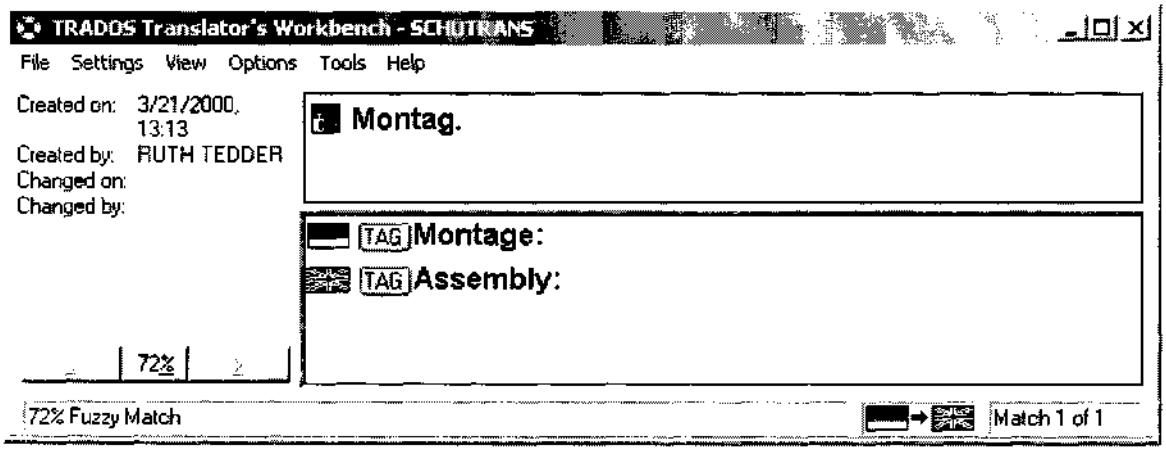
tags. When texts have previously been translated in non-tagged format, the same text appears in tagged format will not find a 100% match. The converse is also true when non-tagged text is translated, and the same applies when the same text appears with different tags, or even when the same tags appear with different text!

a) Finding a match

As has frequently been pointed out in the past, Translation Memory (TM) software is merely an oversized search and replace tool. It doesn't, as many would have us believe "'remember' previously translated sentences"" So when tags are present, they are included in the search and replace operation. And although TRADOS offers the option of reducing the significance of the tags, they can still hamper the translation process. The translation memory is unable in many cases of finding previously translated sentences without tags or with different tags. This can be overcome if translators are encouraged always to use the concordance facility on a portion of the untagged part of the text, even when the segment produces a "no match". It is surprising how often an apparent "no match" has actually already been translated almost in its entirety.

b) When is a match not a match? When it has tags!

Some examples below show what happens in the case of complex formatting resulting in more tag than text. This can happen with headings, subheadings and bullet points, but should serve as a warning to those who would pay translators according to how much of a match already exists (more on this later). In the example below, even without tags, a subheading which happened to be a day of the week: the German *Montag* apparently finds a 72% match when it comes across a different German heading *Montage,* meaning Assembly. Clearly the two are completely unrelated, and in reality this is a 0% match.



The bullet points shown in Figs. 4 and 5 similarly show how the formatting becomes more significant than the words, and the nearest match offered has nothing to do with the segment to be translated. Even though Workbench® will create a penalty for different tags, what it offers as a possible translation will often be wide of the mark.

---

[2] Gilderson, Alan (2000) *Building Blocks Translation Memory* in ISTC Communicator Spring 2000 Vol. 6, No. 9

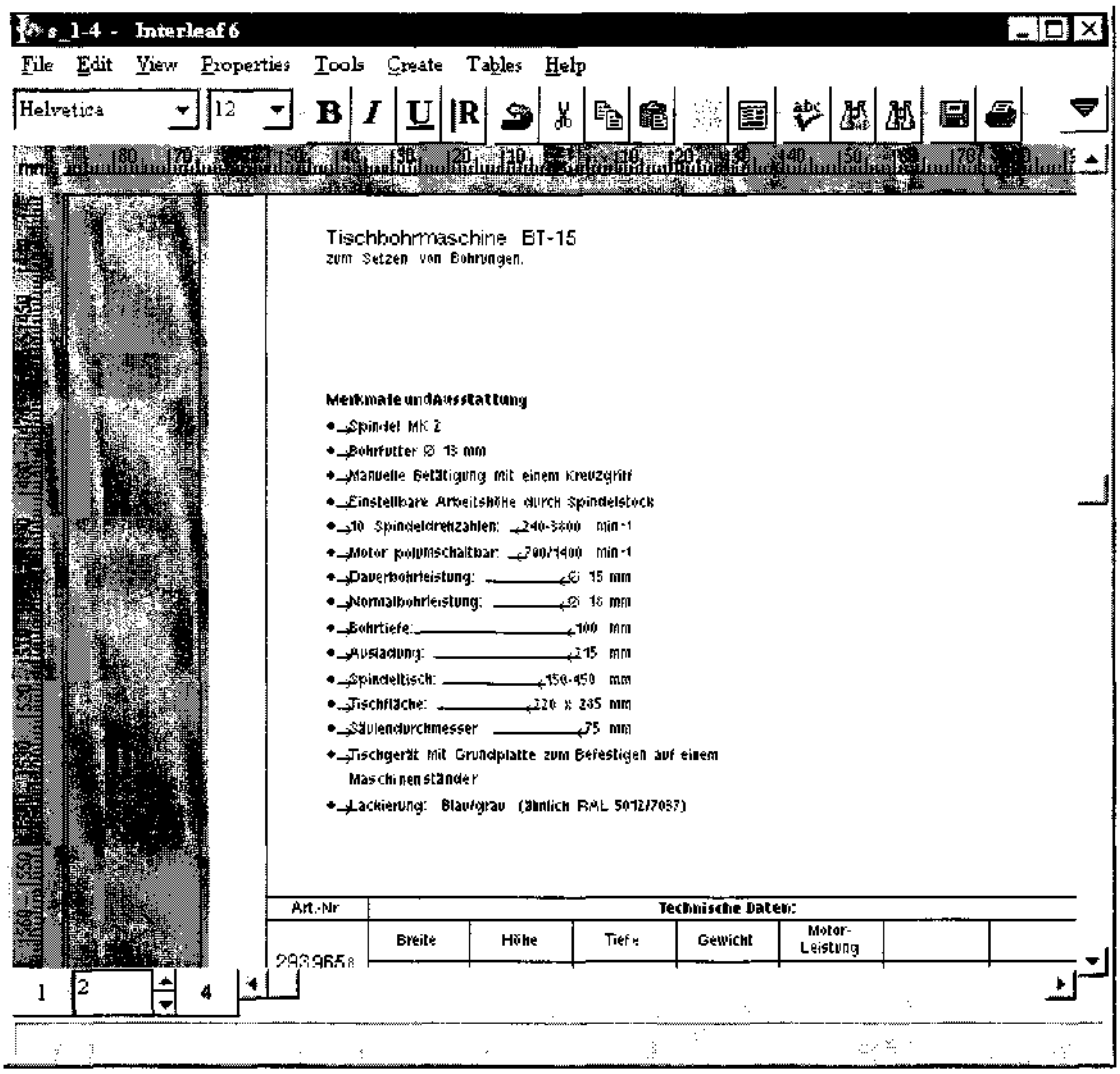


Fig 4 The layout of a text in Interleaf showing formatting such as tabs and bullets

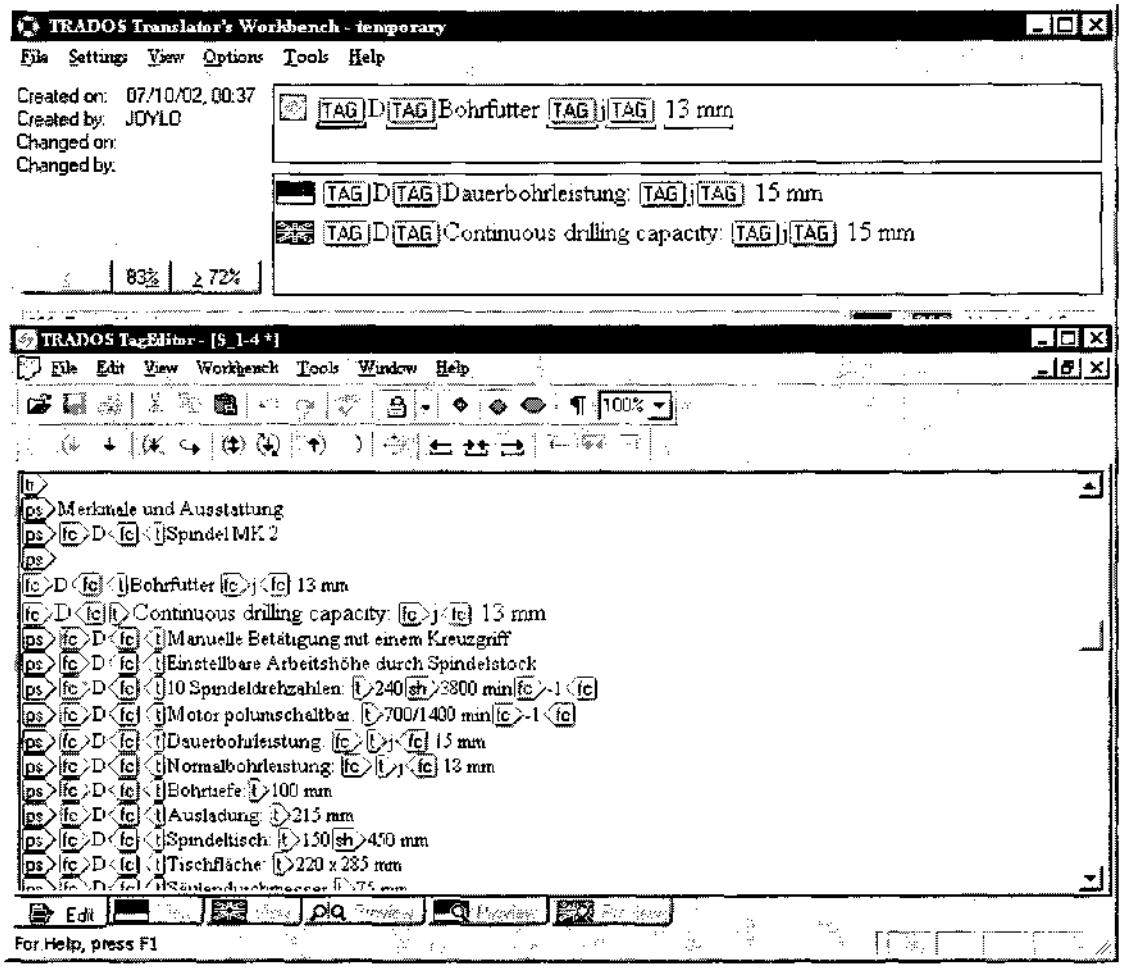


Fig. 5
The same text in tagged format with match found by TM. Deemed to be an 83% match, but in truth 0%.

Another annoying anomaly arises from the problem Trados has with numbers as placables. If a six digit number has a thin space between the third and fourth digit, Trados does not recognise it as a 'placable' (or only recognises the first part of the number) causing an incorrect number in an apparently correct segment. As seen in Fig. 6.

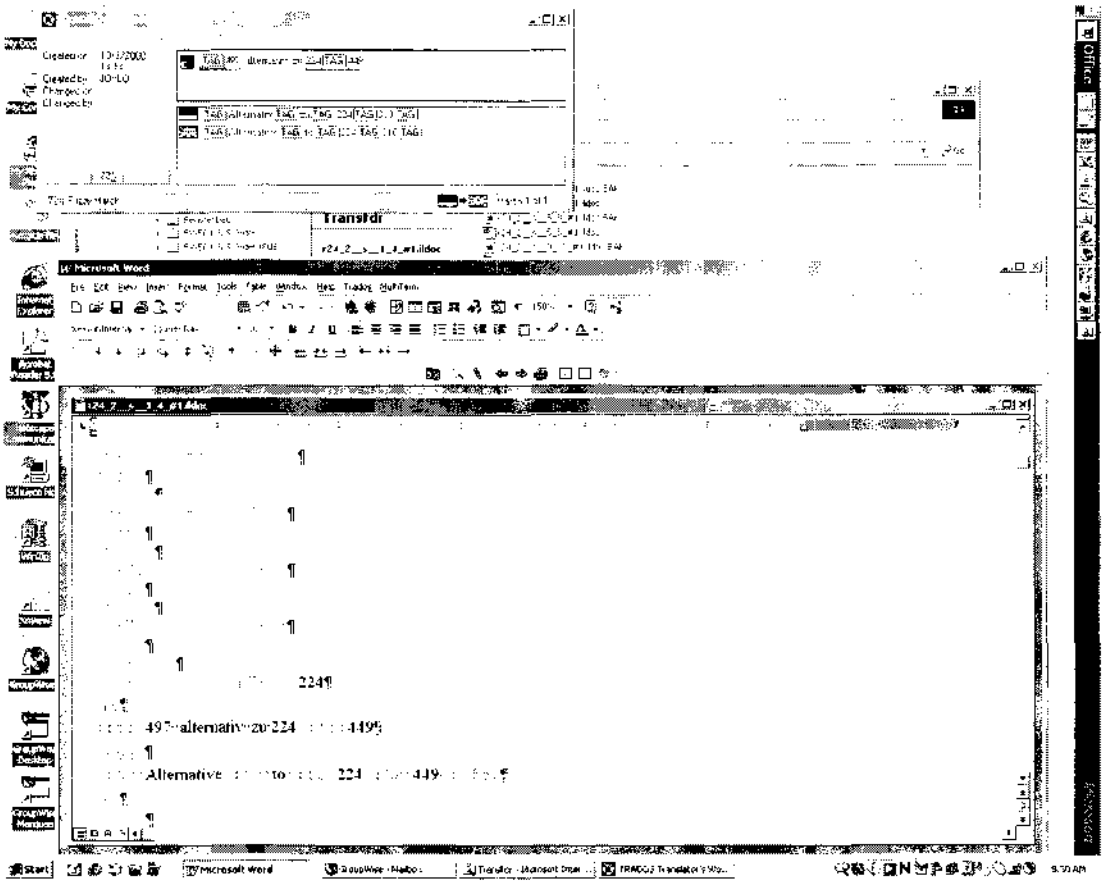


Fig 6. The problem of split numbers

c) <u>If the tags don't match, Workbench® will **try** to help out!</u>
This happens when the text is the same, but the tags are different - perhaps signifying different fonts, bold or italicised words etc. Particularly if the fuzzy match features more tags than the original, the translation memory tool then replaces the tags on a one for one basis, tag for tag, regardless of what the tag is. If there are a number of tags and the translator doesn't notice, this can cause incorrect tags in the final version, which may impede reconversion into the DTP format. This is when the translator stops translating and becomes an editor of tags!

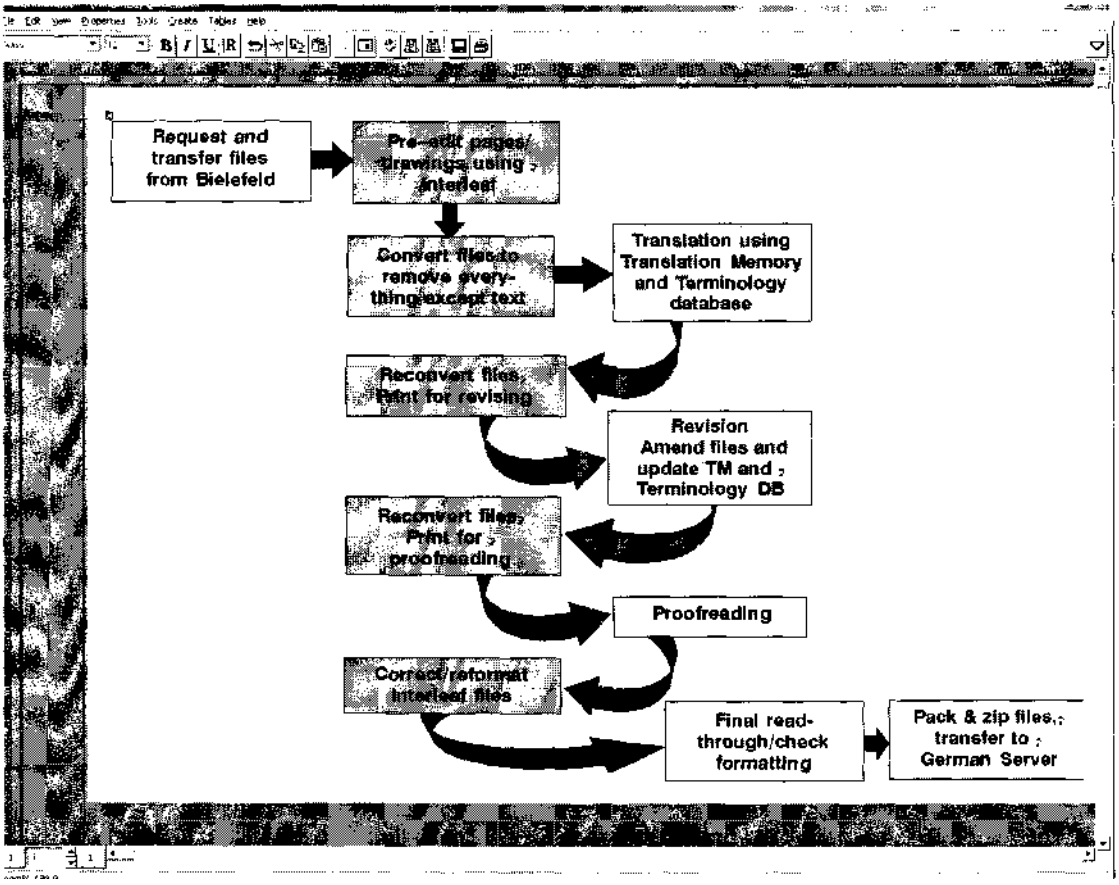
The next problem occurs when the translation in tagged text format is filtered back into the original from whence it came. If the translated text runs longer or shorter than the original, there is the additional problem of ensuring correct layout for publication. Another job that is not really translation.

Case Study; How we avoided tag tyranny

SCHÜCO is a construction engineering company with its head office in Germany and a small in-house translation department in its UK subsidiary to produce English versions of all the technical documentation and sales and marketing material for publication. The majority of the documentation is produced in Germany using Interleaf ®Quicksilver DTP program. The English translation team introduced TRADOS software (Version 2) into the equation in 1998, upgrading to Version 3, and now working with 5.5.
It was noticed early on that poor typesetting in the original documents produced text with multiple tags, many of them serving little or no purpose. Where text was bold

within a normal text component, there were tags, as there were for the frequent use of two different font sizes in the same component. The decision was made initially to save multiple version of the same sentence with different tag formulations, in order to maximise matches in the future. Incorrect formatting, hard returns and tabs, were edited out by the translators within the tagged text file. Much time was expended at reconversion on correcting tag mismatches.

However, as has been discussed at this conference in previous years, if translators are, as they should be, paid to translate, they should not be used to perform tasks which can be done by less qualified or differently qualified personnel. The decision was therefore made to employ a publishing technician to pre-edit the Interleaf files and to carry out all the tasks relating to production for publication. By analysing the different work processes required in producing quality English documentation from the German provided, a workflow was devised with the aim of maximising translators' time spent translating and attempting to maximise the match rate from the Translation Memory. As the illustration in below shows, the publishing technician, in addition to ensuring that the original documents are well formatted, provides a printed version and a tagged text for the translator to work on, reconverts the tagged text after translation and reformats the DTP file for publication.



The translators translate by referring to the printed version to overcome the problems of decontextualisation, and use TM, but in particular using the concordance function, to translate the tagged text segments, large and small and store them for later reuse. The translators also verify the tags, since broken or missing tags mean an incorrect translation unit and this must also be amended in the TM (therefore falling within the translator's domain). When these working methods were first adopted, the translators

in the organisation had previously translated by overwriting the text in the DTP format documents and had learnt to understand the tagged text from their knowledge of the publishing software itself. However, in spite of staff changes and the infrequency of use of the publishing software by translators, the involvement of the translators within the whole translation/publication cycle has resulted in efficient processing of files, with fewer tag errors now the norm.

Maximising the matches when translating tagged text

The following points were found to be useful in maximising matches in an in-house translation department.

1) Always have the fully formatted text to hand when translating
2) Avoid translating mechanistically (see 1 above)
3) Ensure the original is well formatted to reduce extraneous tagging. Even extra time spent pre-editing will produce the pay-off of better matches in future.
4) Allowing a sentence to be stored more than once with different tags attached will increase subsequent 100% matches (but this means more vigilant memory maintenance i.e. ensure that the way the sentences are expressed is not different just because the tags are.)
5) Ensure that external translators are trained in what the tags signify, and can verify translations and repair broken tags.

In short, what is required is a holistic approach to the whole procedure. In spite of the role taken by the publishing technician, the translator cannot translate 'decontextualised' tagged text in isolation, especially when parts of it have been pre-translated. A quality finished job can only be achieved when the translator is completely involved in the project. Where the project is large, translators must be able to co-operate, even if this is in distributed team. Sending a chapter each to 10 different translators who do not communicate with each other will not result in a good finish, unless there is extensive post-translation editing.

Although the workflow shown may not be appropriate to translation agencies which use predominantly free-lance translators, nevertheless, the points raised should alert such organisations to problems which might arise. Similarly, it should give those commissioning translations an insight into the complexities both in the processing of the text and consequently in the translation itself. Non-linguists who commission translations have been known to complain that numbers in the text shouldn't count, as they do not change, even though knowing where the number needs to be placed within the text may be essential to convey the correct meaning. And certainly there seems to be a consensus that in tagged text no payment should be made if only the tags are different, as the tags do not have to be translated. So how much is manipulation worth?

Which brings us back to the debate on how translation is valued, and whether sliding scales of payment based on percentage matches in pre-existing translation memories can be completely justified.

Discussion

This paper has shown that the presence of tags means that "no matches" may be found when some really do exist, and apparent matches with mathematical values of over 80% can hide the fact that none of the words in the segment match in any way at all. Admittedly, the segments in the illustration were not particularly long, but in these days of sound bites and bullet points on websites and presentations, short translation units abound. Even without the tags, there is dubious merit in arguing that a numerical percentage of an existing translation unit is in some way a measure of the work required of the translator in a given circumstance. Perhaps this offers a challenge to the translating profession too. Are translators demeaning themselves by charging by the word? Tags or no tags, a 500 word technical marketing brochure will require more work than an equivalent number of words describing the basic components of a machine.

It has been shown that translation destined for websites and for XML based applications, will in future result in a greater need for testing out the translation within its application to ensure quality, and that it is for translators to test that quality. This cannot be calculated on a word or line basis, but will need to be project based. Computer programmers, after all, are not paid by the number of lines of code they write, so why should such a yardstick be applied to translators, who are certainly no lesser mortals. If translators were paid per project, like a lawyer is paid per case, the problems arising from the use of tools such as those which extract text in various tagged formats, could be tackled internally, as it were, amongst translators themselves. Without the constant argument about how much of a text may or may not have already existed in another form, the aim of a good finished job would stand more chance of being fulfilled.

Conclusion

Whereas some of the tagging problems experienced with older style desktop publishing will no doubt be improved with increasing compatibility between applications, the growing use of platform independent metalanguages like XML means the demand for the translation of "decontextualised" text looks set to increase. The topic of cultural localisation has not been touched on here, but will also need to be included in deliberations. To overcome the problems discussed here, translators need to emphasise the dimensional complexity of translation and insist that they be allowed to offer the quality of which they are capable. Translation customers also stand to gain, not only from time saved on so-called post-editing, but from the quality image they will be able to maintain. No doubt, this debate will rumble on for some time, but ultimately it will be in the interests of both the translation professionals and those commissioning them, and to their mutual advantage to find a joint solution.

References

Culler J.: Saussure Fontana/Collins, Glasgow 1976

Frazer, J.: Part of the team: how to get the best out of translators in ISTC Communicator Vol 6, No. 6, Summer 1999.

Gilderson A.: Building Blocks Translation Memory in ISTC Communicator, Vol 6 No.9, Spring 2000.

Jones I.: Managing a translation service to maximize quality and efficiency, in Translating and the Computer 20. ASLIB, Nov. 1998

Knauf, A.: Development, Use and Profitability of Translation Memory Systems in TC-Forum Vol. 4, December 1999, http://www.tc-forum.org

Myerson C.: Global Economy - Beyond the Hype in Language International, Feb. 2001, www.language-international.com

Rintanen K.& Zetzsche J.: Integrating Translation Tools in Document Creation, 2002, www.translationzone.com

Sklair S.: Multilingual Component Management: Trends and Implications for Translation, in Translating and the Computer 21. ASLIB, Nov. 1999

Zetzsche J.: Translation Databases for Web Site Localization in International Writers Group LLC, www.internationalwriters.com