

Translation Memories guarantee consistency: Truth or fiction?

Joss Moorkens

Centre for Next Generation Localisation

School of Applied Language and Intercultural Studies, Dublin City University, Ireland

www.cngl.ie

Joss.Moorkens2@mail.dcu.ie

Abstract

Translation Memory (TM) is a key technology in the translation industry, its success based on several assumptions: that it saves time, provides cost savings, and enhances consistency. Our research tests the assumption of consistency in TMs. We examine English/German and English/Japanese TM data from two commercial companies with a view to measuring the levels of consistency (or inconsistency) between TMs across product releases.

To meet our aim we first had to develop a method for interrogating TMs for consistency. Following a pilot study, we chose to categorise translation units based on whether the TM translation process had introduced consistency or inconsistency. Additionally, we examined source text segments that contained minor inconsistencies where the corresponding target text introduced further inconsistency. By identifying categories of inconsistency along with investigating the accompanying metadata, this case study hopes to highlight the types of inconsistency that can occur in TM data and to suggest how TM tools might overcome this problem.

Introduction

Use of TM tools has become widespread since their introduction in the early 1990s based on the assumptions that they save time, reduce costs, and enhance consistency. These assumptions are repeated in promotional literature from software producers and translation vendors, as well as in reviews, and articles (Elimam, 2007). As the popularity of these tools have grown, a small number of researchers have noted that consistency may be reduced by translation units (TUs) that have “become inaccurate over time” (Bowker, 2002, p116). Following a pilot study in 2005, Bowker

found that student translators were not critical enough of proposed translation from a TM that had been seeded with errors and wrote that “although it is frequently claimed that TMs improve consistency, this is not always the case” (2005, p18). In 2007, Ribas López opined that dissemination of errors throughout a project from errors in a TM was an underreported phenomenon (2008, p52). This paper aims to provide empirical evidence of inconsistency in TMs, then to categorise these inconsistencies, and to compare results from TMs originating from different sources and in different language pairs.

Typology of inconsistencies

In order for consistency to be identified and measured in TM data, following a pilot study, we developed a typology of consistency for TUs. Within this typology the following four categories are possible:

1. inconsistent source segments are translated as inconsistent target segments
2. inconsistent source segments are translated as consistent target segments
3. consistent source segments are translated as inconsistent target segments
4. consistent source segments are translated as consistent target segments

In category 1 we use the term ‘inconsistent source segments’ to refer to cases where there are very minor formal differences between two source segments and such differences do not reflect any semantic differences between the segments in question. Such minor formal differences include differences in:

capitalisation, tags, punctuation, spaces, and spelling (where a segment may be inconsistent with another segment simply because of a misspelling or a typographical error in one of the segments).

These minor source segment inconsistencies were prevalent in our pilot and main study data and, aside from the typographical errors, would present a 100% match or a 99% fuzzy match to a

translator using the TM.

Thereafter, segment-level inconsistency is observed where two segments that one could reasonably expect to be formally identical differ from each other in some way. In the case of target segments, it appears reasonable to expect segments that are translations of 'the same' source segment (i.e. segments that are translations of different tokens of the same source type) to be formally identical, especially in a translation memory scenario where the goal is to reuse existing translations for already encountered source segments.

Where there are two different translations (and thus two different target segments) for a single source segment type, we speak of target segment inconsistency. The differences between the target segments in question can be very minor formal differences (as defined above), but they can also be more substantial, in extreme cases even leading to semantic differences between the two segments.

Category 2 TUs contain target segment consistency that has been introduced in the translation process. Category 3 was our initial focus for this study – introduced inconsistency in the target segments. Category 4 may be seen as the ideal in localisation, whereby the TM has provided the best possible leverage and thus saved the maximum possible amount of time and money.

Inconsistent segments are counted by identifying the number of types n . The number of segment-level inconsistencies is the type count minus one ($n-1$). Thus in the case of a single source segment (type) that has 4 tokens, if there are 3 separate translations (3 types; one of which appears twice), then the number of target segment inconsistencies is 2 (or $3-1$). Thus we give a special status (of 'master' or 'reference' segment) to one of the target segments, and treat the other two segments as inconsistent with that reference segment. The reference segment is the one which appears first chronologically, and which a translator could have, but did not reuse in unchanged form. For example, the following four translations for 'Click an empty part of the drawing area.' appear in the TM data:

- a. *Klicken Sie auf der freien Zeichenfläche.*
- b. *Klicken Sie auf einen freien Bereich der Zeichenfläche.*
- c. *Klicken Sie auf einen freien Bereich der Zeichenfläche.*
- d. *Klicken Sie auf einen beliebigen freien Bereich auf der Zeichenfläche.*

Although there are four tokens, there are only three types: a, b, and d. If we assign the status of

reference segment to segment a, the segments that are inconsistent with the reference segment are b (repeated for c) and d: thus we count two inconsistencies. When we have three types $n=3$, and since our count is of type $(n - 1)$, we count two inconsistencies.

At segment-level, source or target segments are either consistent or formally differ and are thus inconsistent. However, there may be more than one inconsistency within these segments. For this reason we also count and categorise inconsistencies found within inconsistent target text segments (those found in categories 1 and 3 above).

These inconsistencies are categorised mostly per part of speech aside from those with inconsistent punctuation or where word order has been changed. If there are more than three inconsistencies within a target segment, we consider that segment to have been wholly retranslated. These categorised inconsistencies may be further subcategorised; for example nominal inconsistencies that differ lexically, or in number (singular/plural).

These subsegment-level inconsistencies are counted in the same way as segment-level inconsistencies: we identify the number of types n , assign one the status of master or reference segment, then count the types that are inconsistent with the part-of-speech or word order in the reference segment. Thus the count is n minus the reference segment $(n-1)$. Again, the reference segment is the one which appears first chronologically, and which a translator could have, but did not reuse in unchanged form.

TM data used in this study

In order to maximise the validity of this research, it was important to obtain real-world TM data from several sources (Susam-Sarajeva, 2009, p7). However, obtaining a company's "translation family jewels" (Smith, 2008, p23) proved to be difficult, and these were only received following protracted negotiations and, in one case, after a non-disclosure agreement had been signed. We were further limited to English, German and Japanese by researcher's own language competence. The TM data was presented in the TMX format, and this was parsed using a Python script before the TUs were categorised. The aligned segments were loaded into a LibreOffice spreadsheet for categorisation. While it was possible to automatically highlight inconsistencies of categories 2 or 3, the minor inconsistencies in the source segments for category 1 could only be identified manually.

We obtained four sets of TM data from two companies. TM A is an English to German TM containing 22,691 TUs of aligned segments of which 188 contain only numbers, dates, or punctuation. The remaining 22,503 TUs were categorised as specified above. TM B is an English to Japanese TM from the same project as TM A, containing 18,799 TUs. After removing those that contain only numbers, dates, or punctuation, 18,650 TUs remained to be categorised. TM C is an English to German TM containing 301,583 TUs. After removing those that contain only numbers, dates, or punctuation, 293,924 TUs remained. Due to time constraints, we chose a sample of the first 50,061 TUs to analyse, having confirmed that the incidence of category 2 and 3 TUs in the remainder of the TM was consistent with the sample. This was also the case for TM D, containing 298,700 TUs in English and Japanese from the same project as TM C. After removing the TUs that contain only numbers, dates, or punctuation, we were left with 292,258 TUs of which 50,000 were subject to analysis.

Results by category

Category 1: Inconsistent source text segments

	TM A	TM B	TM C	TM D
Letter case	55	13	370	753
Punctuation	47	2	50	73
Tags	42	2	46	137
Typo	4	1	13	11
Space	10	2	30	68
Word Order	0	0	1	0
Total Segments (tokens)	353	47	947	1980
Total Subsegment Inconsistencies	158	25	510	1042

Table 1: Inconsistencies found in category one source text segments

Category 1 TUs contain minor inconsistencies as specified in our typology in the source segment and other kinds of inconsistencies in the aligned target segment. The number of category 1 TUs found in our four sets of TM data differed, but in all four TMs the most prevalent category of source text (ST) inconsistency was in letter case or capitalisation of words. None of the TT segments aligned with ST segments that contain inconsistencies in letter case themselves contain instances of inconsistent letter case; rather the TT segments in question evince other kinds of inconsistencies, as

in the following example from TM C:

1.1s (*SHIFT+right-click the drawing area.*)

1.1t (*Klicken Sie bei gedrückter UMSCHALTTASTE mit der rechten Maustaste in den Zeichenbereich.*)

1.2s (*Shift+right-click the drawing area.*)

1.2t (*Klicken Sie bei gedrückter UMSCHALTTASTE mit der rechten Maustaste in den Zeichnungsbereich.*)

In both TT segments the ST word 'shift' has been translated as 'Umschalttaste' and capitalised. This would suggest that a TM match was used despite the change of case in the second ST segment. However, the German translation of 'drawing area' was changed from 'Zeichenbereich' to 'Zeichnungsbereich'. According to the metadata, segment 1.1t was created on December 22nd 2006 and last changed two years later on December 7th 2008. Segment 1.2t was created by a different translator on January 15th 2009 and last changed one year later on January 18th 2010.

We found a higher incidence of inconsistent placing of the space character in TM D. These spaces were initially noticed by automatically comparing the ST segment and the following, seemingly identical, ST segment, as 54 of the 68 space inconsistencies were at the end of the segment following a full stop. Again, the aligned target text (TT) segments contain other kinds of inconsistencies, as in the following example:

2.1s {1}lsp{2} file.

2.1t {1}lsp{2} ファイルから自動的にロードされます。

2.2s {1}lsp{2} file. [Contains extra space]

2.2t {1}lsp{2} ファイルは変更しないでください。

The ST segment contains a space inconsistency, while the aligned TT segments differ by particle (は 'wa' and から 'kara'), 2.1t has the additional adjective *automatic* or 自動的 'jidouteki', and the verbs differ semantically and in form. These TT segments also provide examples of explicitation in Japanese TT, a phenomenon that appears throughout TMs B and D. The translation of a ST noun has become a sentence in the Japanese TT, containing detail not present in the ST. Thus we have 'You may load automatically from the lsp file.' in segment 2.1t and 'Please do not change the lsp file.' in segment 2.2t.

Category 1: Inconsistent TT segments

	TM A	TM B	TM C	TM D
--	------	------	------	------

Noun	80	11	330	616
Punctuation	3	4	46	129
Tags	5	4	2	147
Verb	42	2	51	75
Space	6	0	29	141
Word Order	39	1	81	3
Preposition	12	N/A	137	N/A
Particle	N/A	1	N/A	42
Complete (Not added to total)	11	4	9	12
Total Segments (tokens)	353	47	947	1980
Total Subsegment Inconsistencies	225	24	684	1291

Table 2: Inconsistencies found in category one target text segments

In the table above, the number of subsegment inconsistencies may be seen to be larger than that in table 1. This is because a segment may contain up to three subsegment inconsistencies before we consider it completely retranslated. Among category 1 TUs we found a large proportion of noun inconsistencies, comprising between 36% and 48% of the total number. For example, in TM C there were 330 noun inconsistencies of which 12 showed the influence of the source language in one instance but not in another, and 87 contained inconsistencies of capitalisation or case, as per the following example:

3.1s *At the Command prompt, enter subtract.*

3.1t *Geben Sie in der Befehlszeile differenz ein.*

3.2s *At the command prompt, enter subtract.*

3.2t *Geben Sie an der Eingabeaufforderung
DIFFERENZ ein.*

Example 3 also displays a phenomenon that accounts for the high prevalence of preposition inconsistencies in TM C. We found 137 preposition inconsistencies in category 1, just under 20% of the total. 126 of these preposition inconsistencies (and thus 18% of the total) are secondary changes as required by the change of noun, thus we see an alternation between the phrases 'in der Befehlszeile' (in the command line) and 'an der Eingabeaufforderung' (at the command prompt).

Category 2 TUs (Inconsistent ST segments with consistent TT segments)

	TM A	TM B	TM C	TM D
Letter case	140	272	461	505
Space	93	72	118	247
Punctuation	67	4	79	153
Inconsistent word (Noun)	11 (3)	89 (76)	138 (66)	194 (107)
Typo	2	1	19	21
Word Order	1	1	4	0
Total Segments (tokens)	609	854	1667	2183
Total Subsegment Inconsistencies	314	440	811	1123

Table 3: Inconsistencies found in category two segments

Category 2 TUs contain ST inconsistency and thus introduce consistency in the TT. The majority of these ST inconsistencies in all TMs analysed were inconsistent letter case. The following example from TM C is typical of this ST inconsistency.

3.1s *Action Macro*

3t *Aktionsmakro*

3.2s *action macro*

Although capitalisation of the first letter of a German language noun means introduced consistency would be expected in example 3, there are instances of the ST letter case being retained in the TT in all of these TMs (roman lettering is sometimes used in the Japanese TT), particularly if the ST segment is in upper case. This means we have a mix of transposed ST punctuation or formatting and native TT formatting in TT segments. In the following example from the same TM, containing a ST space inconsistency similar to those found in all four sets of TM data, we see a German noun in lower case:

4.1s *{1}securityoptions {2}*

4t *{1}sicherheitsoptionen{2}*

4.2s *{1}securityoptions{2}*

Inconsistent punctuation is usually to do with the presence or absence of commas or full stops in the ST which may or may not be retained in the TT. The following example from TM D contains a

punctuation inconsistency, but also contains an example of a section that has been marked out in the TT, followed by a comment by the translator, explaining that he chose the term 塗り潰し色 'nuritsubushihiro' for filling in colours. This comment was subsequently propagated within the TM.

5.1s {1} If None (Color) is selected as the Map Type then 5t {1}

the color needs to be selected.

[マップの種類として[###塗り潰し色を選択した場合は、色を

5.2s {1} If None (Color)) is selected as the Map Type

選択する必要があります。■3-

then the color needs to be selected.

(B037)「なし」という選択肢はなく、「塗り潰し色」という選

択肢が表示されるので、このようにしました。(Koizumi

06/11/21)

There are a number of reasons why ST inconsistencies may be ignored by a translator who chooses instead to accept a fuzzy match. Foremost among these in Japanese is the presence of plurals in the ST. In our study of English to German TMs, plurals did not register in our categories, as we considered that the ST had formally changed and thus accepted that the TT would be inconsistent. However, as there is no distinction between singular and plural in Japanese – numbers are given explicitly or are implicit in context – we can expect to see plural and singular nouns translated consistently in the Japanese TT, and this is indeed the case. Of 76 cases of inconsistent nouns in the ST segments of category two TUs in TM B, 42 differ in number: singular in one case, plural in another, as in example 6:

6.1s Dimension

6t 寸法¹

6.2s Dimensions

Category 3 TUs (Consistent ST segments with inconsistent TT segments)

	TM A	TM B	TM C	TM D
Noun (SL influence)	84 (18)	48 (9)	197 (31)	365 (77)
Verb	37	36	30	59
Punctuation	12	7	44	183
Space	1	1	82	272

¹ 'Sunpou'.

Explication	1	6	3	32
Word Order	6	4	17	16
Preposition	7	N/A	112	N/A
Particle	N/A	13	N/A	57
Complete (Not added to total)	3	7	9	35
Total Segments (tokens)	408	230	877	1713
Total Subsegment Inconsistencies	185	116	450	1070

Table 4: Inconsistencies found in category three segments

Category 3 contains TUs with inconsistent TT segments, where inconsistency has been introduced in the TM data. Again, the most prevalent category of TT inconsistency was noun inconsistency. In TM A we found 84 inconsistently translated nouns (46% of the inconsistencies) of which 18 showed influence of the English source language in one instance as in example 7:

7s All lines that have been converted using the {1}Create surface borders{2} function can be recognized easily since they are drawn with the {3}Border{4} pen.

7.1t Alle Linien, die mit der Funktion {1}Flächenränder anlegen{2} konvertiert wurden, können Sie leicht erkennen, da sie mit dem Stift mit der Bezeichnung {3}Border{4} gezeichnet werden.

7.2t Alle Linien, die mit der Funktion {1} Flächenränder anlegen{2} konvertiert wurden, können Sie leicht erkennen, da sie mit dem Stift mit der Bezeichnung {3}Rand{4} gezeichnet werden.

This alternation between 'Border' and 'Rand' occurred three times in the TT and was one of several patterns that emerged within the data.

The Japanese TT in TM B again showed detail being added in translation that was not in the ST. We found ten inconsistencies that were translations of the word 'selecting', as shown in example 8:

8s *Selecting*

- 8.1t エレメントの選択
- 8.2t コールアウトエレメントの選択
- 8.3t アセンブリの選択
- 8.4t 多角形の選択
- 8.5t 線の選択

- 8.6t 楕円の選択
- 8.7t 選択
- 8.8t 長方形の選択
- 8.9t ベジエ曲線の選択
- 8.10t めねじの選択
- 8.11t おねじの選択

In the example above 8.1t is taken as the reference translation as it appeared first in the TM data. Each TT segment contains the noun 選択 (*sentaku*) meaning 'selection' but most involve further explicitation. 8.1t is エLEMENTの選択 or 'selection of elements'. Segment 8.2t is コールアウトELEMENTの選択 or 'selection of callout elements'.² While this explicitation may make the TT segments clear and understandable, it has a negative effect on leverage. It may be in this case that the first translation contained added detail that was not appropriate for the subsequent translations.

After noun inconsistencies, the next most prevalent category in TM B is verb inconsistencies. Of the 36 verb inconsistencies, 18 of them contained another repeated pattern, alternating between using the verb 拘束する (*kousoku suru*) meaning 'to bind or restrict' in one case, and the verb 関連付ける (*kanren tsukeru*) meaning 'to relate' in another. For example:

9s Binding ISO Elements to XML Elements using Object Info 9.1t オブジェクト情報を 使用して ISO エLEMENTを XML エLEMENTに拘束する

9.2t オブジェクト情報を 使用して ISO エLEMENTを XML エLEMENTに関連付ける

Segment 9.1t translates as 'bind ISO elements to XML elements using object information', segment 9.2t as 'relate ISO elements to XML elements using object information'. Looking through the metadata, each verb choice is not attributable to a single user ID, but the translations using 拘束する were all saved to the TM at the same time on April 22nd 2009. Two uses of 関連付ける were also saved then, but all others were dated from the 7th of May in 2009. At that stage, one would presume, the TM tool used must have suggested the previously translated TT segment as a 100% match.

The other Japanese data, TM D, also contain a repeated pattern, alternating between the borrowed

² They continue with アセンブリの選択(8.3t): 'selection of assembly'; 多角形の選択(8.4t): 'selection of polygon'; 線の選択(8.5t): 'selection of a line'; 楕円の選択(8.6t): 'selection of ellipse'; 選択(8.7t): 'selection'; 長方形の選択(8.8t): 'selection of rectangle'; ベジエ曲線の選択(8.9t): 'selection of Bezier curve'; めねじの選択(8.10t): 'selection of female screw'; おねじの選択(8.11t): 'selection of male screw'.

English word レイヤ and the native Japanese word 画層 'gasou' 41 times as per example 10:

10s *A new layer group filter can be nested only under another group filter.*

10.1t 新しいレイヤグループ フィルタは、他のグループ フィルタに対してのみネストできます。

10.2t 新しい画層グループ フィルタは、他のグループ フィルタに対してのみネストできます。

TM D contains a large number of punctuation inconsistencies. Many of these (23) are marked out using the # symbol, others have inconsistently placed quotation marks, and many show indecision as to whether or not to retain ST formatting for commas or full stops as in the following example:

11s *Accesses Dimensioning mode*

11.1t 寸法記入モードにします

11.2t 寸法記入モードにします。

The high rate of preposition inconsistency in TM C is again a secondary effect of noun inconsistency as shown previously in example 3. The inconsistencies of particle in Japanese are also often secondary to a change in verb or verb form from active to passive or, as in the following example, required by verb choice with the 表現 (scale representation) taking the particle 'ga' when the verb 'aru' (to exist) is used, and the direct object particle 'wo' with 'motsu' (to hold).

12s *Annotative objects may have multiple {1}scale representations{2}.*

12.1t

異尺度対応オブジェクトには複数の{1}尺度表現{2}があります。

12.2t

異尺度対応オブジェクトには、複数の{1}尺度表現{2}を持つものもあります。

Category 4 TUs (Consistent ST segments with consistent TT segments)

These TUs are those that we consider to have been translated consistently. By looking at the number of repeated TUs that fall into this category, we can see the overall rate of introduced TT inconsistency within a TM as per the following table.

	TM A	TM B	TM C	TM D
Category 3 TUs	408	230	877	1713
Category 4 TUs	6674	4302	18343	25541
Total TUs with repeated ST segments	7082	4532	19220	27254
Percentage of TUs with introduced inconsistency	5.8%	5.1%	4.6%	6.28%

Conclusion

Our initial aim for this study was to find a satisfactory method for interrogating TM data. While we feel that the categories as specified in our typology have provided us with detail of the content of the data, the manual nature of separating particularly category 1 TUs was time-consuming. In this study we focussed particularly on TT inconsistency. In categories 1 and 3 noun inconsistency is prevalent across all TMs. These inconsistently translated nouns vary between borrowed words, influenced by the ST, and native words in both German and Japanese. Often two nouns from different time periods or as chosen by different translators will have been propagated throughout the TMs, causing inconsistency in repeated patterns.

Inconsistent use of spaces and punctuation appears to have a negative effect on leverage within TMs, and spaces at the end of segments, while not visible to a writer or translator, may reduce the match percentage, introducing the possibility of inconsistency. Inconsistent ST punctuation adds to a lack of clarity in these data with regard to textuality (Torres-Hostench et al., 2010, p260): translators appear to be undecided as to whether to retain ST formatting or to change to suit the TT, so both are left in the TM to be propagated.

As shown in example 8, explicitating insertions make a segment less useful for leverage and appear to have the knock-on effect of suggesting changing the insertion in a fuzzy match rather than translating the ST segment as presented. Insertion of comments also poses a danger to leverage, as these have also been propagated through TMs. While these would be better written into the metadata, perhaps there is an issue with regard to interoperability between tools and retention of comments and messages between translators. However, as translators are often not permitted or motivated to edit exact matches, inconsistent data is likely to be propagated within TMs and used in

translations.

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.engl.ie) at Dublin City University.

Future Work

This quantitative study is intended to form part of a sequential mixed methods project on inconsistency in TMs. Within the coming months we intend to carry out a series of interviews with translators and translation reviewers based on the initial research and their experiences of consistency in TM data. It is hoped that this will add detail to the study and explain some of the findings in more depth.

References

Bowker, L. 2002. *Computer-aided translation technology: A practical introduction*. Ottawa: University of Ottawa Press.

Bowker, L. 2005. "Productivity vs quality? A pilot study on the impact of translation memory systems". *Localisation Focus* 4(1): 13-20.

Erimam, A. S. 2007. "The impact of translation memory tools on the translation profession". *Translation Journal* 11(1). [Accessed from <http://translationjournal.net/journal/39TM.htm>]

Ribas López, C. R. 2007. *Translation Memories As Vehicles For Error Propagation: A Pilot Study*. Minor Dissertation. Tarragona: Universitat Rovira i Virgili.

Smith, R. 2008. "Your Own Memory". *The Linguist* 47 (1): 22-23.

Susam-Sarajeva, S. 2009. "The case study research method in translation studies". *The Interpreter and Translation Trainer*. Manchester: St Jerome. 37-56.

Torres-Hostench, O., Biau Gil, J. R, Leal, P. C., Mor, A. M., Mesa-Lao, B., Orozco, M., Sánchez-Gijón, P. 2010. "TRACE: Measuring the impact of CAT tools on translated texts". In Gea-Valor, M., Garcia-Izquierdo, I., Esteve, M. (eds.) *Linguistic and translation studies in scientific communication*. Berlin: Peter

Lang. 255-276.