LISA

2002 Translation Memory Survey

Translation Memory and Translation Memory Standards

http://www.lisa.org

The *2002 LISA Translation Memory Survey: Translation Memory and Translation Memory Standards* was written by Arle Lommel (LISA). The survey form was designed and created by Alex Lam (LISA) with significant input from LISA's OSCAR Special Interest Group and others on the LISA staff. LISA and the author would like to thank OSCAR for its valuable input on the survey, as well as all those who participated in the survey.

Table of Contents

Executive Summary

The 2002 LISA Translation Memory survey was conducted on-line at the Localization Industry Standards Association (LISA) website from August through October 2002. 134 respondents answered questions about their usage of Translation Memory (TM) technology, about the role of Open Standards such as Translation Memory eXchange (TMX), and about their interest in the exchange of TM assets between companies.

The survey respondents represented a mixture of job descriptions from freelance translators to the C-level, with the vast majority coming from the managerial, C-level and GILT Professional categories, as shown in **Figure 1. Survey Respondents**.

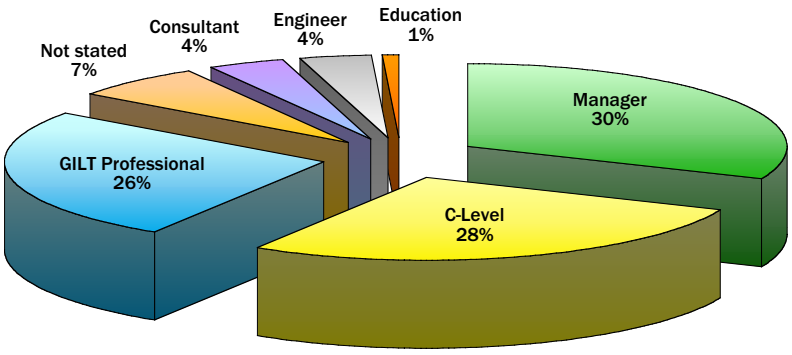

***Figure 1. Survey Respondents**. Job titles of survey respondents in broad categories.*

Companies represented ranged in size from those translating less than 1 million words a year to those translating in excess of 50 million words per year, with a fairly even mix of translation volumes.

Significant findings

The survey found that use of TM technology among those surveyed is rapidly maturing, with 87% having used it for at least one year. The majority of respondents use TM for everything they can and are looking at extending TM usage to areas they have not been able to use TM for yet. Respondents also showed considerable sophistication in their understanding of the limitations and strengths of TM software.

TM users display sophistication in use of technology

Additional findings include the following:

- Early adopters (those having used TM more than five years) tend to have the largest TM asset repositories – roughly 15 times the size of those using TM for 1–5 years and in excess of 400 times the size of those using TM less than one year. TM repositories represent a significant investment and strategic asset for companies and late adopters of TM will be unable to achieve the productivity gains of the early adopters until they have had time to build their TM assets.

Time makes the difference in TM asset size

- Contrary to expectation and preliminary findings, translation volume is not significant in determining how long companies have used TM, but the longer a company has used TM, the greater the amount of its translation work that will be done using TM. This may represent a greater awareness of authoring and linguistic issues on the part of companies that chose to adopt TM early on.

Translation volume does *not* predict degree of TM usage

29% of companies already use TMX

- Almost 90% of respondents have TM assets in commercial tool formats, 29% use TMX, and 11% use internal formats.

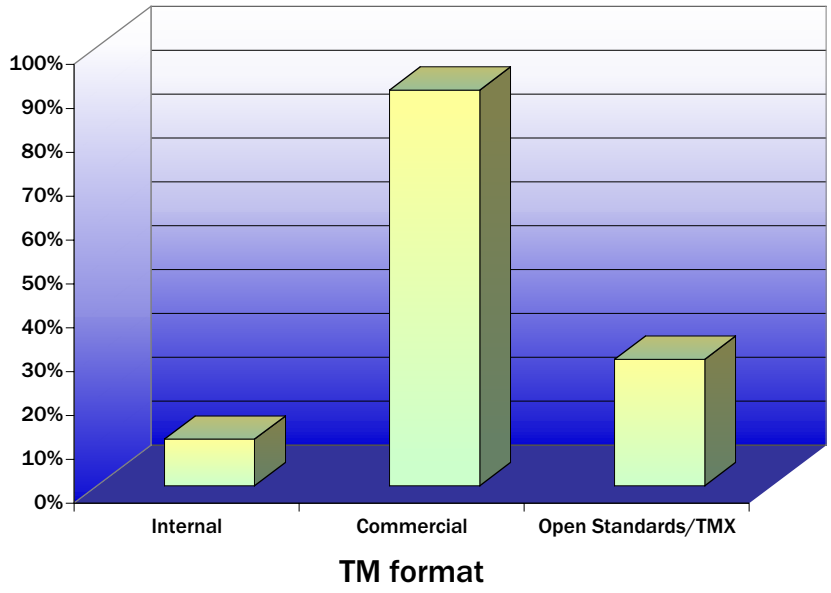

Figure 2. Formats in which Translation Memories are Stored. *Bars show percentage responding in each category.*

A few sophisticated users account for a large percentage of TM

- Users of internal and commercial or open-source tools have TM assets larger than other users, by an order of magnitude. These users, although a small minority of all users, account for a large portion of all TM assets held by companies

29% report client requests for TMX

- The most-cited reasons for adoption of Open-TM standards (like TMX) are exchange purposes and support for open standards. 29% of respondents indicated client requests as a motivation for the adoption of TM standards. Ensuring reusability and easing integration of additional technologies are also important drivers for adoption of standards.

Almost 90% use their TMs for purposes beyond "simple" translation

- The most common operation performed on TM segments is QA, followed by term replacement in segments. Given the prevalence of these tasks, many users would benefit from improvements in tools for these tasks. Almost 90% of respondents report using TM segments for purposes beyond "simple" translation.

- Half of TM users are interested in exchanging TM assets with other companies, while a third have no interest or cannot exchange TM for legal or quality reasons.

Conclusions

User sophistication and understanding of TM has increased considerably in recent years, as has the market coverage of TM tools. Since the bulk of users have been using TM technology for more than one year it seems likely that TM tool developers cannot rely solely on market growth for increased sales. Increasingly users will expect and demand more from their TM systems, leading to increasing competition among developers on the basis of features. As the size of TM assets increases the need and demand for TMX and other standards related to translation memory will grow.

Survey Respondents

134 GILT professionals responded to the survey, available on the LISA website from August 9, 2002 through October 9, 2002. Most of the respondents were from localization service providers or translations departments, with some free-lance translators and representatives of academic institutions. Respondents provided their job titles as part of the survey. Most of the respondents fell in the categories *Manager*, *C-Level*, and *GILT Professional*. The broad job categories were determined as follows:

- *Manager* includes job titles like "Project Manager", "Coordinator", "(Department) Manager", and "Account Manager". (31%)
- *C-Level*, in addition to job titles specifically referencing the C-level (e.g., "CEO", "CTO"), includes "Owner", "Director", "Managing Director", etc. (28%)
- *GILT Professional* included translators, writers, and others working in localization not included in any other category. (26%)
- *Engineering* includes job titles specifically referencing engineering. (4%)
- *Consultant* includes job titles specifically referencing consulting. (4%)
- *Education* includes the job title "Professor". (1%)

An additional 7% of those surveyed did not identify their job position. (Numbers do not add up to 100% due to rounding errors). In some cases job titles were ambiguous, especially between the *Manager* and *C-Level* categories, but the majority of respondents are clearly in supervisory positions within their companies.

Translation Volumes

Survey participants were asked to indicate the translation volume of their company in millions of words. Participants represented every volume level, with especially large representation from the ends of the spectrum: the largest blocks were *>50 million source words* (23%) and *<1 million source words* (19%).

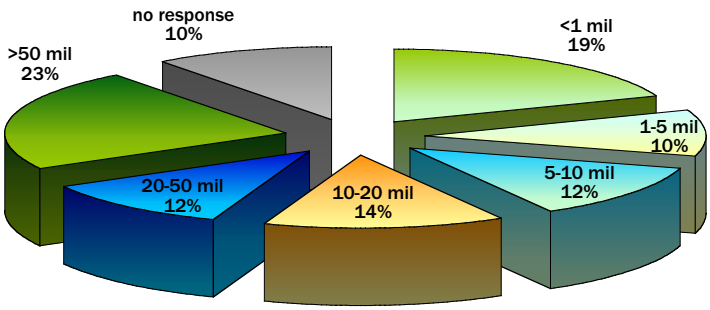

Figure 3. Translation Volume in Millions of Source Words per Year.

10% of those surveyed did not indicate their source word volume, possibly because in some cases this was not known to those answering the survey or because they did not have figures for the company as a whole.

Translation Memory Usage and Trends

Most companies using TM use it for most of their translation

Virtually all of those surveyed indicated that they presently are using Translation Memory tools. Most of those surveyed use TM for the majority of their translation work, as shown in **Figure 4. Use of TM**:

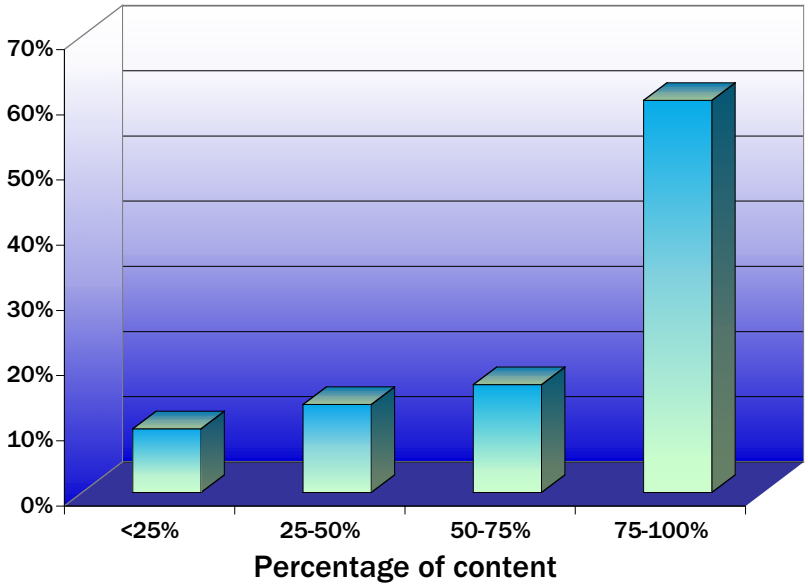

Figure 4. Use of TM. *Percentage of content processed using TM. Almost 60% of those surveyed indicated that they use TM for 75%–100% of their translation work.*

Translation volume does not predict levels of TM usage

Most users of TM use it for most of their translation, regardless of translation volume, as shown in **Figure 5. TM Usage vs. Translation Volume**.

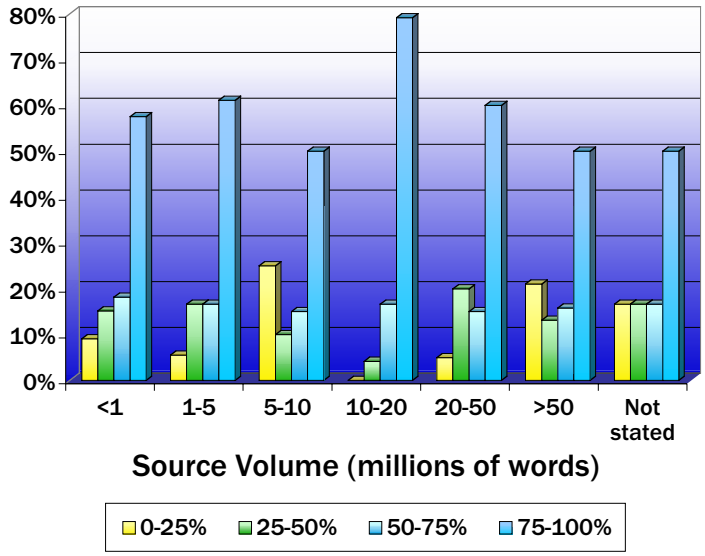

Figure 5. TM Usage vs. Translation Volume. *Percentage of content processed with TM at various levels of translation volume.*

Surprisingly there is no discernible trend for companies with high translation volume to use TM for more of their content than those with lower translation volumes. This finding

contradicts preliminary findings that had indicated that companies with high translation volumes tended to use TM for more of their translation requirements. At all volumes at least half of the respondents used TM for 75–100% of their translation.

When asked if they intended to expand their use of TM beyond its present limits slightly more than half indicated that they did intend to.

Most companies either use TM for everything possible, or plan on extending usage

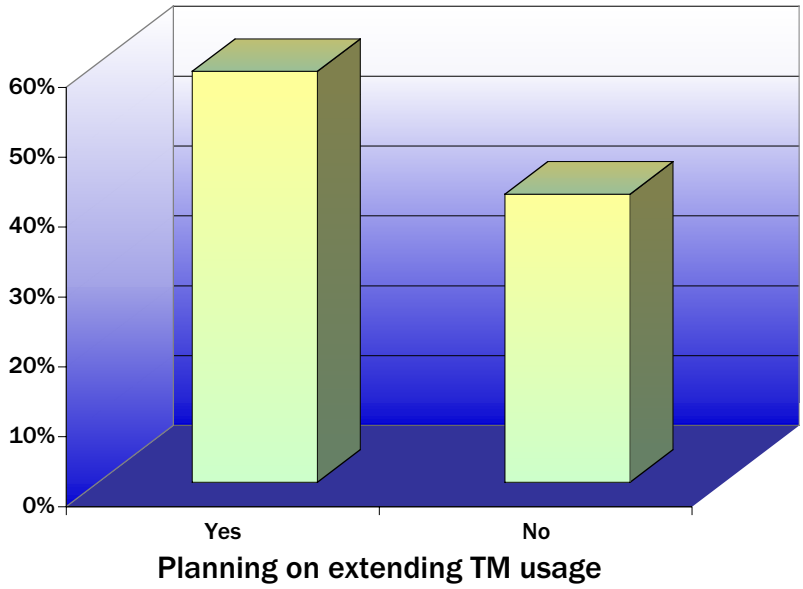

Figure 6. Extension of TM Usage.

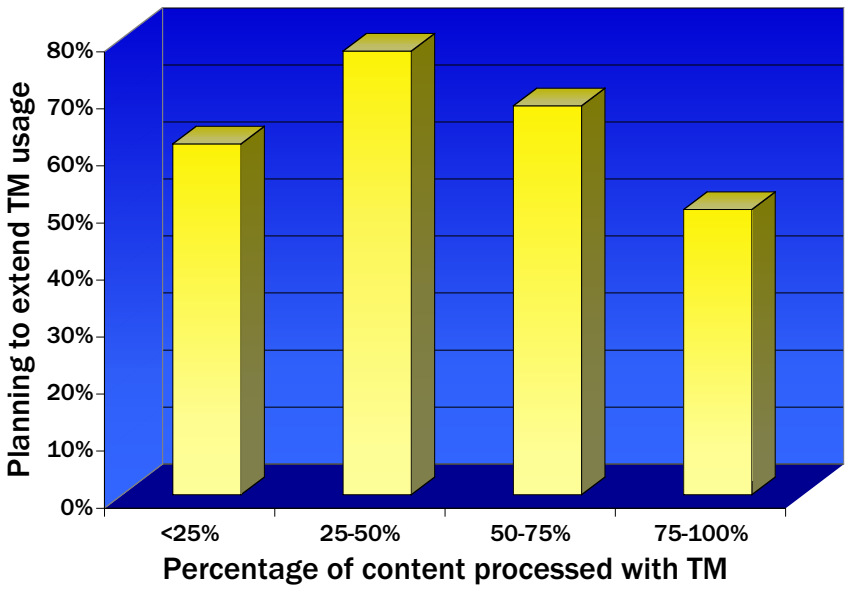

Figure 7. Extension of TM Usage vs. Present TM Usage.

At all usage levels except the 75–100% usage level, the majority of those surveyed intended to extend usage beyond present levels. However at least half of those in the 75–100% usage category not planning to extend usage, are either using it for all content, or deal with content types not amenable to TM processing. Most of those commenting on reasons for not extending TM usage show real awareness of the limitations of present TM systems with regards to file types and text types.

Most TM users have used it more that one year

Most of those using TM have used it for more than a year (88%) with 32% having used it for more than 5 years. 11% have used it for less than a year, as shown in **Figure 8. Length of TM Usage**.

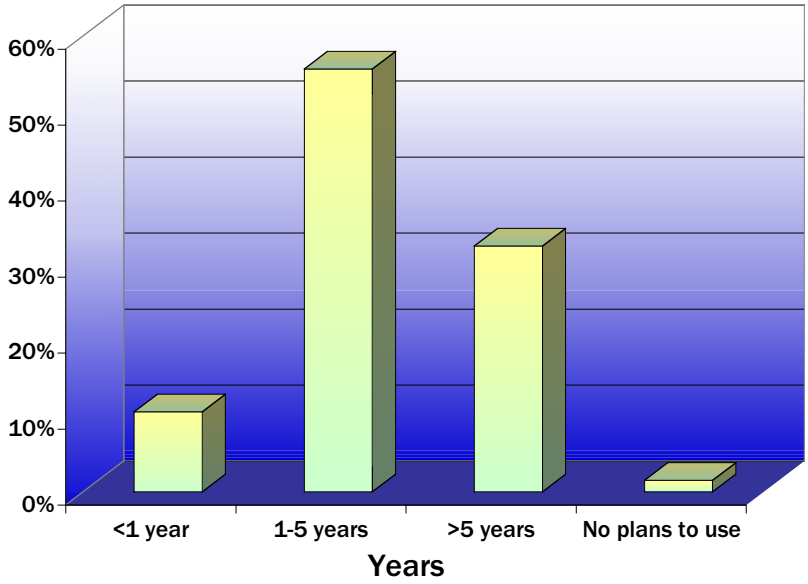

Figure 8. Length of TM Usage.

The longer a company has used TM, the more it will use it

Although translation volume does not correlate to usage levels of TM, length of TM usage does. The longer a company has used TM, the more it translates using TM. 79% of those using TM for more than five years use it for 75–100% of their content versus none using it for less than 25%. (By comparison only 23% of those using TM for less than a year use it for more than 75% of their translation.)

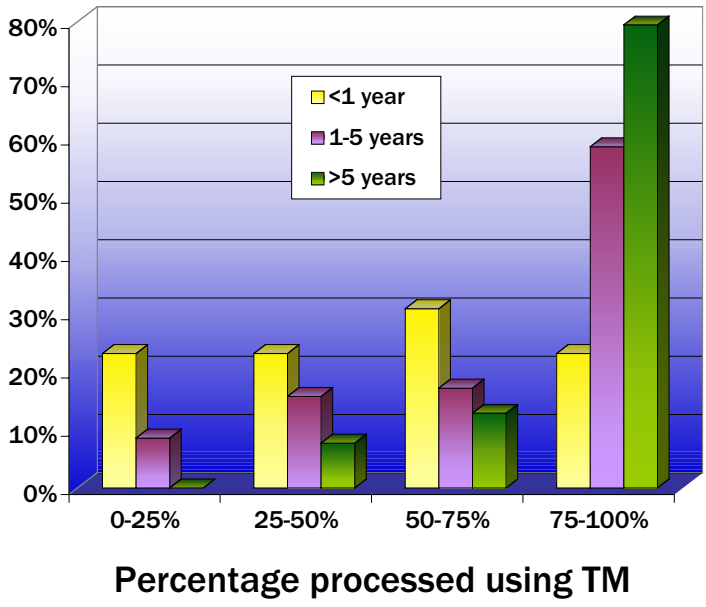

Figure 9. Percentage of Material Translated Using TM vs. Time Using TM. *Percentages represent percentage within each time category.*

When comments concerning reasons why TM is not used for all content are considered it is apparent that many users do not see value in TM use for one-time jobs, small jobs, or certain text types. There is little tendency to see TM systems as one-size-fits-all technology solutions, but rather to use them with discretion where appropriate. Additional reasons for not using TM tools for all work at present include lack of file-format support (PDF appears to be the file format users most wish to add to their TM tools), lack of platform support, and lack of repeated content.

TM Repository Sizes

The longer a company has used TM the more TM it has

Total reported TM assets held by those surveyed averaged 3.42 million segments. As with percentage of content processed using TM, size of TM repositories shows a strong correlation with how long a company has been using TM tools. As shown in **Figure 10. Segments in TM Respositories**, companies using TM for more than five years have literally millions of TM segments in their repositories, while those using TM for less time average in the tens of thousands (<1 year) to the hundreds of thousands (1–5 years).

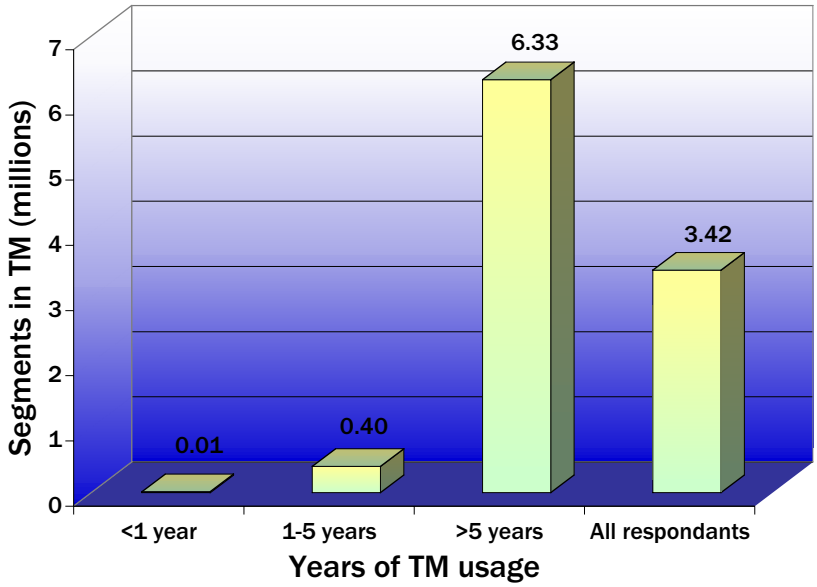

Figure 10. Segments in TM Repositories. *Millions of segments in TM repositories versus length of TM usage.*

This strong time effect should not be surprising, given the cumulative nature of TM assets. Companies using TM tools longer will have had more time to build TM assets than those just beginning to use TM.

Participants were also asked to state how many *files* they have TM stored in. Results from this question were compromised by the fact that differing TM systems store data differently. Most TM programs combine multiple source files or projects into a single database, while others (e.g., Star Transit) store a single TM file for each language pair generated from a source file and combine these on the fly. Thus answers ranged from 1 to 500,000. Using a somewhat arbitrary upper boundary of 500 files for TM tools that generate single database files yields an average number of 55 TM files for single-database TM tools, and approximately 62,000 for Transit-type TM tools. Number of TM files does not correlate with number of segments in TMs.

Functions Performed on Translation Memories

TMs are used for more than just translation

TMs generally require maintenance of various sorts or may be used for quality assurance. What TMs are used for besides "straight" translation is shown in **Figure 11. Functions Performed on Translation Memories**.

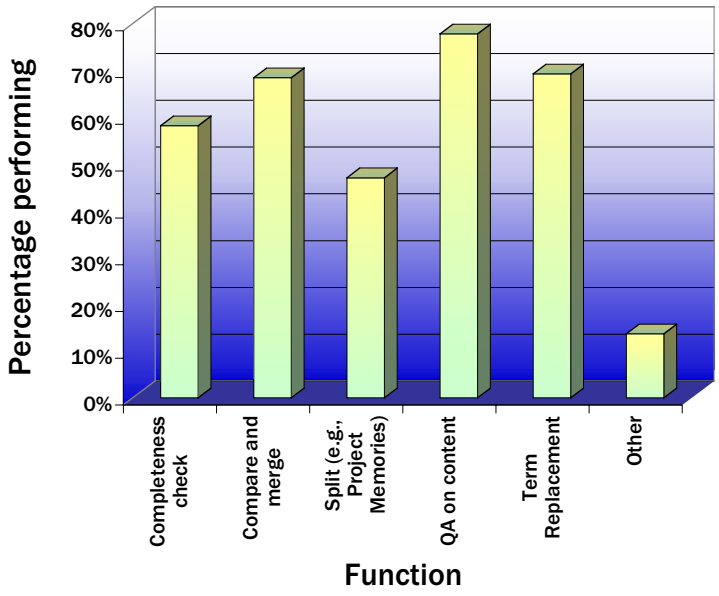

Figure 11. Functions Performed on Translation Memories. *Percentage of those responding who perform various functions on their TM assets.*

Quality Assurance (QA) on content was the most common function, with 78%, followed by *term replacement* (69%) and *compare and merge* (68%). Completeness check was indicated by 58% and Split *(e.g., Project Memories)* received 47%. *Other* received 14%

The "other" category may in fact be under-reported as it relied on survey participants to indicate what they do with their TMs. Other tasks included the following:

- workload assessments
- ongoing QA of *new* material
- backup
- elimination of duplicates
- check against corruption
- correction of numbers, tags and punctuation
- cleaning for re-use in SGML files
- pre-project word counts
- combining translations from various sources
- consistency checking
- term extraction

In particular *term extraction* and *pre-project word counts* are both very common tasks that are likely under-reported in this survey.

Translation Memory Formats and Standards

In addition to providing a snapshot of the scope of companies' usage levels of Translation Memory and how companies are using it, the LISA Translation Memory survey examined the formats TMs are stored in and the impact of TM surveys such as Translation Memory eXchange (TMX), the standard for TM interchange created by LISA's OSCAR.

Translation Memory Formats

The LISA Translation Memory Survey did not attempt to determine usage levels for specific products or their formats. It looked instead at usage levels for three broad categories of TM products/formats: *internal proprietary* (i.e., in-house), *product propriety* (e.g., Trados, Star, WordFast, etc.), and *TMX/open standards*. (For the sake of convenience the product proprietary category will be referred to as *commercial*, regardless of whether a given product is in fact commercial or free/open source.)

90% use commercial tools

As can be seen in **Figure 12. Formats in which Translation Memories are Stored**, 90% of those surveyed make use of commercial TM products versus 11% using internal proprietary formats. 29% use TMX to some extent.

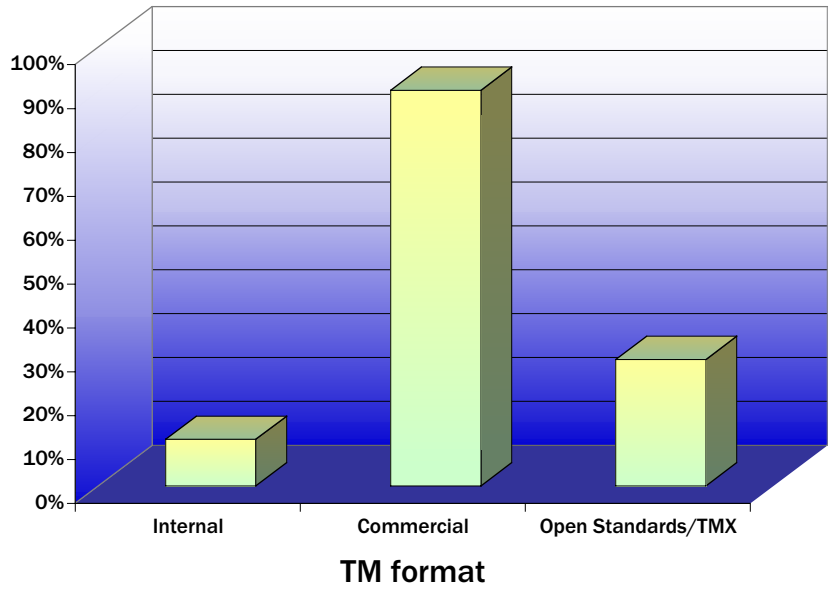

Figure 12. Formats in which Translation Memories are Stored.

This picture is not complete however, since it does not take the volumes of TMs stored in each format. Although users of internal TM formats account for only 11% of those surveyed, the size of their TM assets are four times the size of the average TM repository reported in this survey. Repositories making use of TMX tend to be smaller than either commercial or internal formats. TM repository sizes in each of these formats are given in **Figure 13. TM Size vs. Format**.

Users of internal tools have much larger TM assets than other TM users

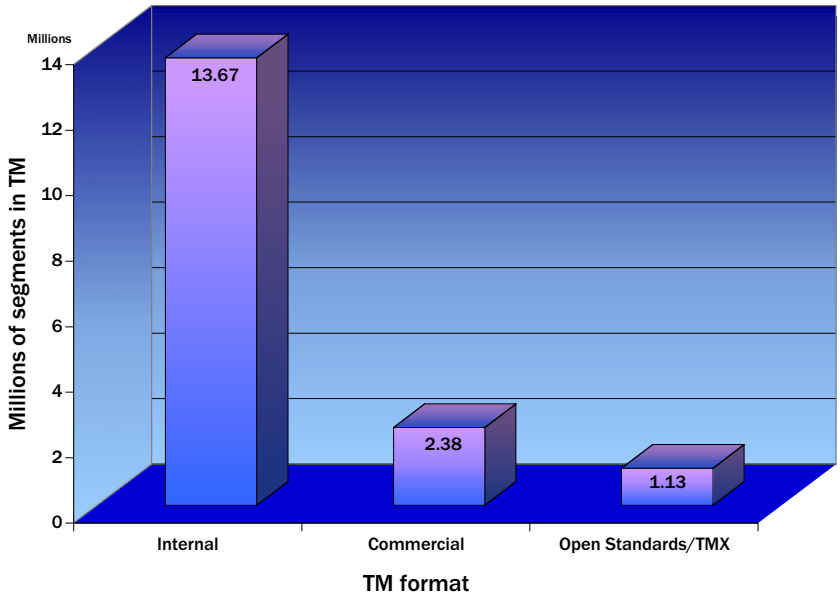

Figure 13. TM Size vs. Format. *Average TM repository size for users of each format.*

In addition it is revealing to break down the information on use of formats to show what combinations of formats are used. **Figure 14. Translation Memory Formats – Complex View** shows these combinations.

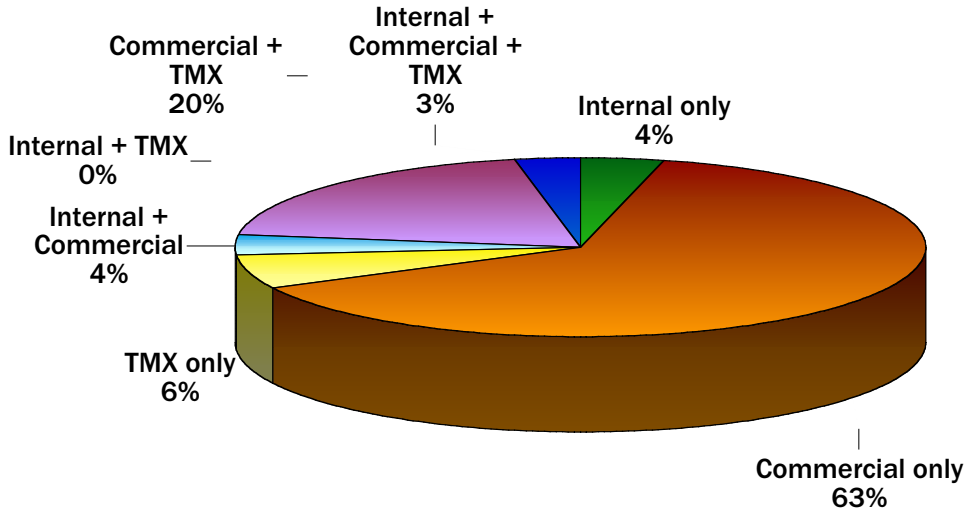

Figure 14. Translation Memory Formats – Complex View.

Users of only commercial tools account for 53% of users

63% use only commercial TM tools. The next largest group are users of commercial TM tools plus TMX at 20%. Surprisingly 6% of respondents indicated that they use TMX only—it may be that these users maintain TM repositories with TMX files generated by external vendors since, at present, it is difficult to conceive how TMs could be useful in TMX only except for archive purposes. (It is also possible that some of these result from errors in filling out the survey.)

Comparison of this more complex view of format usage with translation memory volume provides the results shown in **Figure 15. TM Size vs. Format – Complex View**.

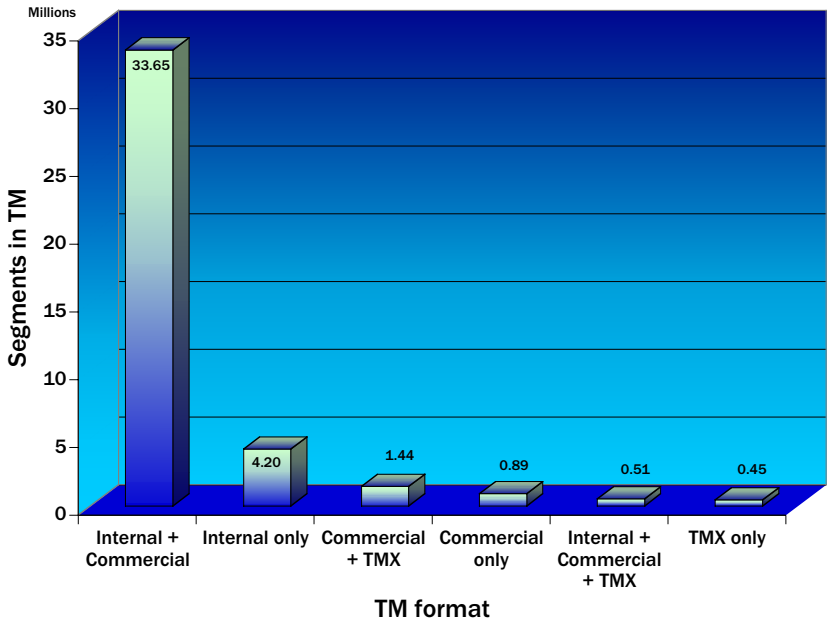

Figure 15. TM Size vs. Format - Complex View. *Average TM repository size for users of various combinations of TM formats.*

In this view it is clear that the primary drivers for the large TM volumes of users of internal products shown in **Figure 13** are companies using commercial tools in combination with internal tools. These users constitute a small percentage of all users (7%), but have, on average, over 21 times as many segments in their TMs than do other categories.

Users with both internal and commercial systems have the largest assets

Adding together the number of TM segments reported in each category gives a quite different view of the importance of formats than does the average, as shown in **Figure 16. TM Formats vs. Absolute TM Volume**.

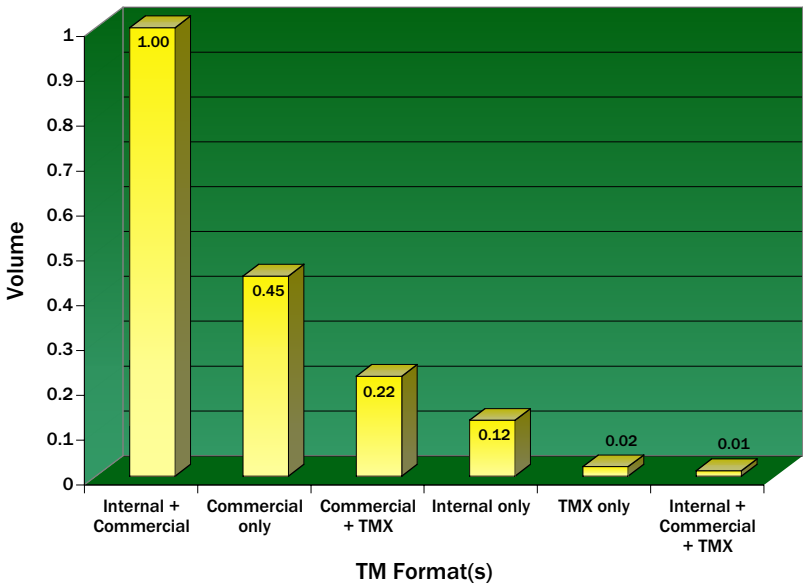

Figure 16. TM Formats vs. Absolute TM Volume. *Total segments in TM respositories reported for each format, adjusted on a scale of 0 to 1.*

(For ease of comparison, all numbers in **Figure 16** were rendered on a scale of 0 to 1.)

In this view combinations of internal and commercial tools is still the largest (and probably most economically significant), but that commercial tools have moved up significantly—despite the much smaller average size of TM repositories stored in commercial tools alone the number of such repositories makes them very important.

Commercial TM systems are likely more important in terms of TM volume than internal systems

The survey did not seek to determine what percentage of TM repositories stored in mixed formats was stored in each format, thus it is impossible to state with any degree of certainty what percentage of the internal/commercial combination is really commercial. However, if the assets in these combinations stored in internal formats are similar in size to those stored in internal formats alone (with the remainder stored in commercial formats), then commercial tools would show up as over 1.3 on the scale used in **Figure 16**, while internal formats would rise to .24. These figures are, however, conjecture and it is impossible to say with any degree of surety how many segments are in internal formats versus commercial formats. Nonetheless it is certain that users of a combination of internal and commercial tools tend to have substantially larger assets than those of any other group.

Users of a combination of tools typically have used TM longer than others

As shown in **Figure 17. TM Format(s) vs. Time Using TM**, users of internal tools tend to have used TM longer than other TM users and those using a combination of internal and commercial tools lead in length of use, with none of them using TM for less than one year.

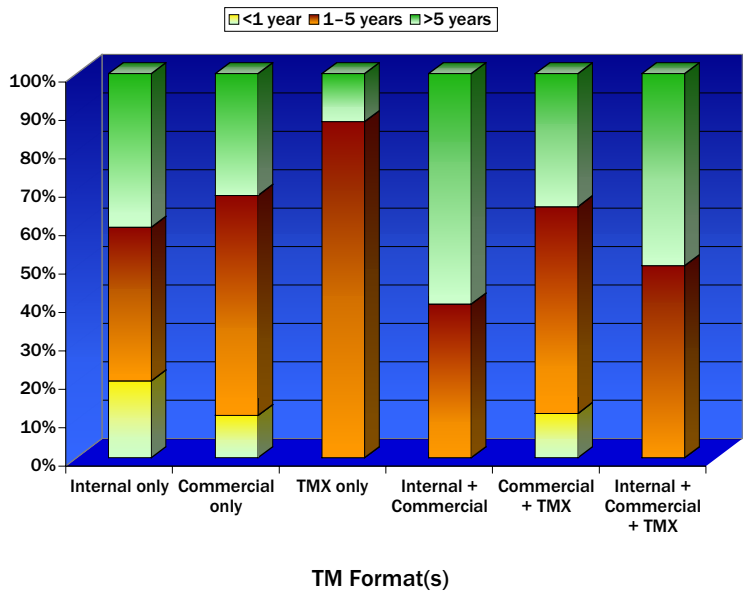

Figure 17. TM Format(s) vs, Time Using TM.

No users indicated that they used TMX in combination with internal tools who did not also use commercial tools. This may be because users of only internal tools find little need to exchange TM data between tools. Relatively few reported using all three format types, perhaps because users of internal and commercial formats may already have found custom data exchange solutions for the tools they use. In addition companies with large assets may take considerable time to evaluate new data exchange solutions or may be waiting for the development of more robust TMX tools that can be integrated into their workflows.

Reasons for Using Open Translation Memory Standards (e.g., TMX)

Figure 18. Reasons for Using TMX shows reasons those surveyed cited why their companies are using or would use open Translation Memory standards such asTMX.

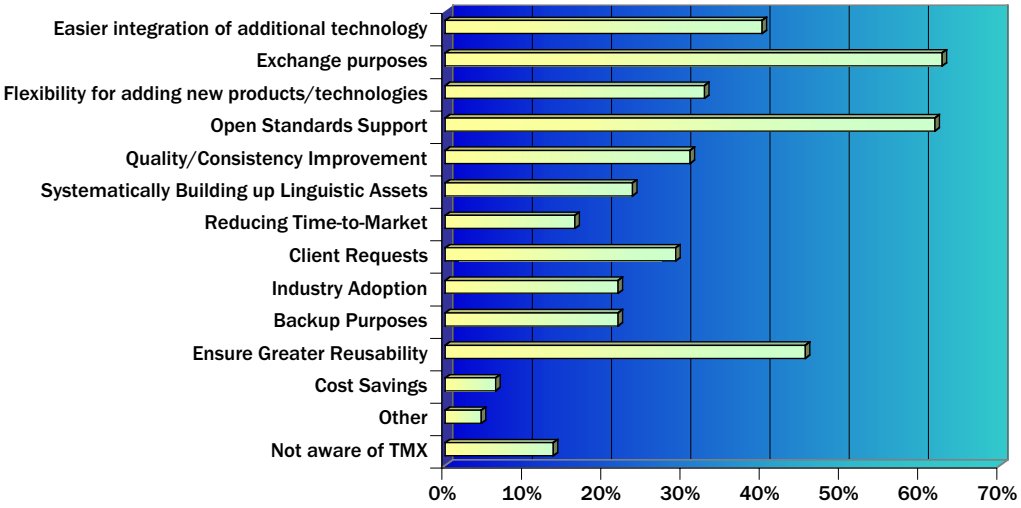

Figure 18. Reasons for Using TMX.

Companies look to TMX to exchange TM data and integrate new technologies

The most significant reasons cited are exchange purposes and support for open standards. Other important reasons are ensuring greater reusability (of TM assets), and easier integration of additional technology. Some tools vendors have indicated that they are not receiving customer requests for better TMX implementation; the results of this survey, however, indicate that demand for TMX does exist, even if this is not being expressed to tools vendors—29% of those surveyed indicated that clients have requested TMX, so vendors are already facing demand for TMX.

Most companies do not look to TMX directly for cost savings, but many of the cited reasons would potentially yield cost savings in certain circumstances.

Although a number of those surveyed cited *other* as a reason for adopting TMX, none of these specified what other uses were a factor.

A question in the survey about what sorts of information would be useful in implementing TMX was intended for internal use by LISA's OSCAR group and its results are not presented here. However, it is apparent that implementation guidelines and business cases would be especially useful in helping companies build a case for use of TMX.

The industry needs to be educated about TMX

Interest in TMX is widespread, but not universal, at this point—14% of those participating in the survey were presently aware of TMX, and so would not have considered reasons for using it prior to the survey. The industry needs education on the uses and benefits of TMX before companies will demand TMX support or make greater use of present support.

Comments on this question indicate that some users of TMX are aware of present difficulties in implementation of TMX and are looking for improvements in TMX applications before making greater use of TMX.

Buying, Selling and Trading Translation Memories

At present there is little, if any, market for the inter-company exchange of translation memory assets (outside of client-vendor relations). Discussion of the idea of buying, selling or trading TM assets has gained recent publicity. Interest in exchange of TM assets on the marketplace is shown in **Figure 19. Interest in Buying, Selling and Exchanging TM**.

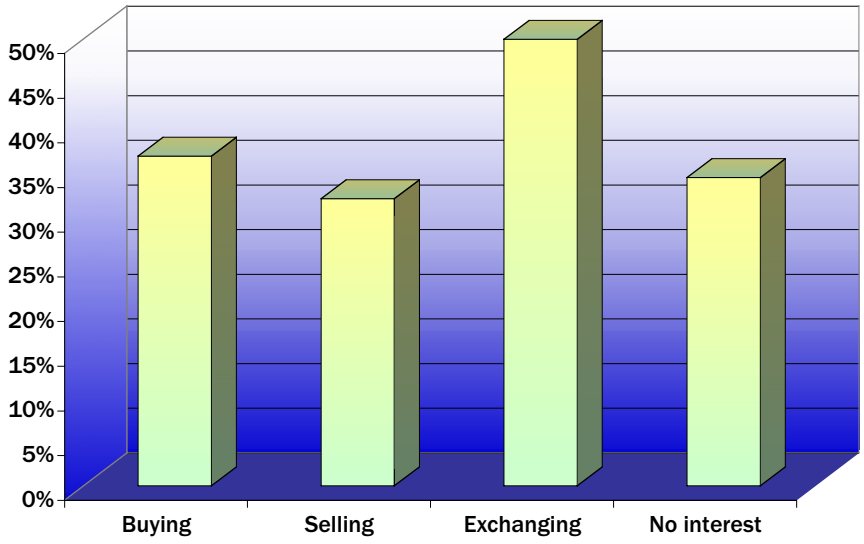

Figure 19. Interest in Buying, Selling and Exchanging TM.

As shown interest in buying TM assets from other companies is at 37% of those surveyed. 32% are interested in selling their own assets. Interest in exchanging assets is at 50%, while 35% have no interest in any exchange of TM assets with other companies at all.

At present the notion of a TM asset marketplace is problematic and those surveyed recognize this, with many citing Intellectual Property and copyright concerns as reasons why they would not exchange TM. Others cited concerns about quality in memories from external sources or said that they would only accept memories developed from materials specific to the same client as their present project. Some respondents cited value in using existing TMs for terminology work, an area likely less problematic than use of third-party TMs directly in translation.

Given that most contracts specify that clients own the memories developed for their projects it would be impossible for vendors to sell memories without client permission, something most clients would be unlikely to grant given the competitive market value of translations in the world market.

Clearly the establishment of a viable marketplace for translation memories will require solution of present problems, yet some of those surveyed indicated that they felt that the exchange of TM assets will eventually be a huge market; interest on the part of some is quite keen.

Appendix A: Survey Questions

Organization Name: _____
Given Name: _____
Family name: _____
Job Title: _____
E-mail address: _____

*Note : Your e-mail will only be used to inform you when the survey results is released. LISA respects your privacy and the information provided will only be used for the above stated purpose.

1. How many source words a year is your company translating ?
   - ❑ less than 1 million
   - ❑ 1–5 million
   - ❑ 5–10 million
   - ❑ 10–20 million
   - ❑ 20–50 million
   - ❑ More than 50 million

2a. What percentage of your content is processed using a TM system of some sort?
   - ❑ less than 25%
   - ❑ 25–50%
   - ❑ 50–75%
   - ❑ 75–100%

   Please list the reasons why you do not use TM for all contents: _____

2b. Do you plan to extend the usage of TM technology?
   - ❑ Yes
   - ❑ No

   If yes, which areas and text types are your planning to extend to ?

3. How many years has your company adopted Translation Memory ?
   - ❑ Less than 1 year
   - ❑ 1–5 years
   - ❑ More than 5 years
   - ❑ My organization has no plans to implement Translation Memory as a strategy

4. How many Translation Memory segments/records does your company currently have?
   - Please provide estimate _____
   - How many files/DBs do you have? _____
   - ❑ My organization does not use Translation Memory

5. What format(s) are your Translation Memory stored in ?
   - ❑ Open Standards (Translation Memory eXchange (TMX))
   - ❑ Internal Proprietary
   - ❑ Product Proprietary (Trados, SDL, Star, Catalyst, ForeignDesk, Systran, DejaVu, TranSuite, Alchemy, Wordfast, etc.)

6. What are your reasons for adopting open Translation Memory Standards (e.g. TMX)? (Check all that apply)
   - ❑ Easier integration of additional technology
   - ❑ Exchange Purposes
   - ❑ Flexibility for adding new products/ technologies
   - ❑ Open Standards Support
   - ❑ Quality/Consistency Improvement
   - ❑ Building up systematically linguistic assets
   - ❑ Reducing time-to-market
   - ❑ Client requests

- ❑ Industry adoption
- ❑ Ensure greater reusability
- ❑ Cost Savings. Estimated savings per annum: _____ (USD/EUR)
- ❑ Other : Please Specify _____
- ❑ I am not aware of the Open Translation Memory Standards such as TMX
- ❑ Backup purposes
- ❑ Independence from TM vendors

7. What will help you or your organization in implementing TMX ?

	Very Useful	Sufficient for now	Not Applicable
Articles on implementing TMX			
Company case studies			
Valuation metrics of Translation Memory assets			
Intellectual Property, Copyright & Legal Considerations for TMX			
TMX ROI Studies & Metrics			
Other useful information: Please specify a. _____ b. _____			

8. Do you perform any operations on your TM segments ?
   - ❑ Completeness check with source material sent to translation
   - ❑ Split (e.g. Project Memories)
   - ❑ Term Replacement
   - ❑ Compare and merge
   - ❑ QA on content (e.g. right language, right terminology) and details
   - ❑ Other (Please Specify) _____

9. Would you trade (exchange, buy or sell) Translation Memory resources ?
   - ❑ I would consider buying
   - ❑ I would consider exchanging
   - ❑ I am not interested
   - ❑ I would consider selling
   - ❑ I will need to investigate further

Comments: _____

The Localization Industry Standards Association
7 Route du Monastère • CH-1173 Féchy • Switzerland
Tel: +41 (21) 821 32 10 • Fax: +41 (21) 821 32 19 • lisa@lisa.org