CONTENT PROFILING AND TRANSLATION SCENARIOS

Sheila Castilho and Sharon O'Brien

Dublin City University

Abstract

Today's companies are overwhelmed with the need to create a huge amount of content, faster, customized, and for numerous media platforms, in order to support their products. Struggling with managing this amount of information, companies have now realised that the strategic management of multilingual enterprise content has become essential. Strategic management involves profiling content, its uses, its end readers and deciding what should be translated, into which languages, using which translation processes and technology. Profiling enterprise content is necessary in order to maximize the quality of the content and its translation at minimum effort and cost by reducing complexity. By targeting the audience, content could be categorized according to the expectation of the end-users, and so, different translation scenarios can be applied to different content types. This article will discuss the challenges of profiling content within the enterprise, as well as translation scenarios focusing on the decisions that push content in one or another direction.

Keywords: Content Profiling, Translation Strategy, Translation, Machine Translation

1 Introduction

Today's companies are overwhelmed with the need to create a huge amount of content, faster, customised, and for numerous media platforms, in order to support their products. According to Boiko (2005), defining content is relevant since organizations have a very simplistic idea of what content is, often confusing data and content.

> Content [...] is a compromise between the usefulness of data and the richness of information. Content is rich information that you wrap in simple data. The data that surround the information (metadata) is a simplified version of the context and meaning of the information. (Boiko 2005, 12)

This misunderstanding between what data and content are may lead to more problems than solutions when trying to deal with content management. Boiko affirms that content is information that was given a "usable form intended for one or more purposes" and its value is "based upon the combination of its primary usable form, along with its application, accessibility, usage, usefulness, brand recognition, and uniqueness" (2005, 8).

Struggling with such large amounts of information, companies now realise that the strategic management of multilingual enterprise content has become essential. Of course, many companies have been engaged in high-volume multilingual content management for decades, but the explosion in content creation and its translation following Web 2.0 has made the management of such content much more demanding. Consequently, several Enterprise Content Management (ECM) systems have been developed in the past few years in order to tackle the problem of content management. While there is no consensus on a suitable definition for ECM systems, Smith and McKeen's definition is widely accepted:

> Enterprise content management (ECM) is an integrated approach to managing all of an organization's information including paper documents, data, reports, web pages, and digital assets. ECM includes the strategies, tools, processes, and skills an organization needs to manage its information assets over their lifecycle. (Smith and McKeen 2003, 647)

As such, ECM systems help companies to keep track of their content by capturing, organizing (indexing, classifying, linking content and metadata) and keeping it up to date. However, although ECM systems enable the indexing and classification of content, they do not directly address the issue of content profiling for translation purposes. Strategic management involves profiling content, its uses, its end readers and deciding what should be translated, into which languages, using which translation processes and technology.

Literature on ECM systems has largely appeared in recent years, but the literature on content profiling is not as well-developed. Rockley and Cooper (2012) is perhaps the best known work on enterprise content strategies in which the authors present detailed information and advice on how to manage different content types within the enterprise. Although the authors mention translation, they do not discuss it in any great detail. Thus, while they affirm that "getting content out to the right customer at the right time and the right format is critical to an organization's success" (Rockley and Cooper 2012, 3), they do not specify how translation processes and technology can contribute to this dynamic.

Profiling enterprise content is necessary in order to maximize the quality of the content and its translation at minimum effort and cost by reducing complexity. However, this is not an easy task. The difficulty in profiling content may be due to the fact that the creation of content is generally not centralized, which causes so-called 'silos':

> Content is created by authors working in isolation from others within the organization. Walls are erected between content areas and even within content areas. This leads to content being created, and recreated, and recreated often with changes or differences introduced at each iteration.
>
> (Rockley and Cooper 2012, 5)

Each time content is created and recreated, the cost and effort increase exponentially. When translation is added to the process, the complexity, effort and cost of translation escalates. "Content silos result in increased costs, decreased productivity, reduced quality, ineffective content, and unhappy customers. The effects of content silos are numerous, costly, and insidious" (Rockley and Cooper 2012, 6).

The issue of inconsistent or poor source language content is mentioned frequently by translators who have to make sense of ambiguous source language content and terminological or stylistic inconsistencies. Of course, the translation of repeated source language content has been catered for by the introduction of translation memory tools. Yet, TMs do not eradicate source language content issues, and can even store them for replication over many translation iterations (see discussion in Moorkens 2012). Poor and inconsistent source language content also contributes to poor quality Machine Translation (MT) output, which increases in turn the post-editing effort.

Recent efforts by the Translation Automation User Society (TAUS) consider the role of content profiling as a precursor to translation quality assessment. TAUS has developed a Dynamic Quality Framework (DQF) (O'Brien et al., 2011), in which they state that quality should be considered prior to translation rather than trying to handle problems with quality after translation. The DQF includes a source content profiling tool, which allows users to categorise their content according to pre-defined categories and according to the channel of communication (e.g. Business-to-Consumer) and the most important communicative functions of the content (e.g. the content should be accurate and clear, the content should engage the reader emotionally). This profiling exercise then results in a recommended model for quality assessment of the translated content. Although still in its early stages, this initative at least attempts to link quality assessment of translated content with the source content profile; these two are unfortunately frequently divorced from each other.

TAUS proposes that guidelines for source creation and translation should be used within the enterprise. They suggest ten meta-categories for profiling content which were elicited from a survey with enterprises. These categories are:

- Audio/Video Content;
- Marketing Material;
- Online Help;
- Social Media;
- Training Material;
- User Documentation;
- User Interface Text;
- Website Content;
- Legal Content;
- Knowledge Base

According to this report, although some companies have some specific content types that do not fit into any of the categories, and some companies do not produce content for all of the categories, they agreed that, in general, their content could be profiled according to these ten meta-categories. It should be noted that the companies who took part in this survey are largely IT multinationals who are actively engaged in ‘localisation’ and so the list above, while relevant to such companies, does not claim to cover all content types from all domains.

Although not mentioning translation per se, Rockley and Cooper also advise that companies should consider the following questions during content creation: “Who needs and uses what content (what content needs to be created, for whom and by whom); How effectively the content currently supports the customer; How content is currently created, managed and delivered" (2012, 10).

By targeting the audience, content could be categorized according to the expectation of the end-users. Questions such as ‘Who is going to read the content?; For what purposes?; In what part of the world?’, can establish important variables. By profiling content, the enterprise can also define how content is going to be created and how it will be translated, if at all.

Both the TAUS initiative and Rockley and Cooper (2012) emphasise the importance of evaluating the (source) content prior to translation on the basis that companies need to know if the content is meeting customers’ expectations. Three parameters were offered in the TAUS Content Profiling system to assist with profiling according to three parameters, namely: ‘utility’, ‘time’; and ‘sentiment’. ‘Utility’ refers to the relative importance of the function of the content (e.g. if it is instructional in nature, it presumably needs to be very clear and consistent); ‘time’ refers to the speed with which the translation needs to be produced (e.g. is it very urgent and needs to be published within 24 hours, or will it be published in several month’s time?); and ‘sentiment’ refers to the relative sensitivity of the text for the brand it represents (e.g. a mistake might be very harmful to the company that produces the text). While this approach makes some attempt to understand source content and how it influences the translated content, it needs to be developed with specific reference to recent advances in machine translation technology, which is on the increase in the localisation sector. In fact, a recent survey (DePalma et al. 2013) on the current state of the language outsourcing localisation market suggests that more companies are adopting automatic translation systems in order to translate enterprise content.

Generating data from the responses of over 1,000 suppliers in the language outsourcing market, DePalma et al. report on the percentage of LSPs that offer a given service or technology such as Translation (Human), Machine Translation Post-editing (MTPE), Translation Technology (which includes CAT tools) and others. According to this report, since 2011 the number of LSPs who offer MTPE has grown from 37.75% to 44.09%. HT went from 94.33% to 96.80, while Translation Technology went from 33.02% to 40.88%.

In the same year, DePalma and Sargent (2013) presented a report based on buyers of language services and MT technology via 108 respondents who use MT in their companies. They found that 88% of those companies have used MT for 1-10 years and the most cited reasons for using MT are: reducing cost; the need for speed; the desire to enter more markets; and the desire to provide better support to international customers. Reasons for not using MT include: linguistic quality, technical complexity, pricing models, lack of language support, etc.

The authors also asked the participants how they see the quality level of MT systems. One percent (1%) said that the quality is ‘excellent’; 10% said it is ‘good’; 66% ‘fair’; 14% ‘poor’; 3% ‘horrible’ and 6% says ‘it depends’. Sixty percent (60%) of the companies publish their MT output after some external or internal post-editing. Only 8% of the companies publish their MT output immediately. In general, MT output is rarely published without some kind of PE. When asked who they target with the MT output content, the participants mentioned the following external audiences - customers (62%), website visitors (40%), and prospects (11%); and internal - for employees.

This discussion hopefully demonstrates that, at least in some sectors, more and more content is being produced and translated, that translation is becoming increasingly technologised, but that content creation sometimes happens within silos and without much consideration for the translated language audience needs. Therefore, content creation and translation is becoming more and more complex. We argue that by profiling content in terms of end-user needs, more informed decisions could be made about what needs to be translated, using which translation processes (human translation, translation using CAT tools, machine translation). Our starting point was to find out how multinational companies with localisation needs are currently profiling content and how translation decisions are made, based on this profiling. To address this question six key decision makers in six companies were interviewed.

2 Methodology for Data Collection

The participants kindly accepted the invitation to be interviewed and the interviews took place between June and August, 2013, in different places according to the suitability and availability of the participants. The interviews were recorded and then transcribed and coded, and a copy was subsequently sent to each participant for review. The interviews consisted of a questionnaire, which the interviewer used to keep track of the questions, but free dialogue was allowed during the interview. Each interview took around 1 – 1:30hrs.

2.1 Companies

The six professionals who accepted to be interviewed were from the following companies (in alphabetical order): Adobe, Autodesk, McAfee, Microsoft, Oracle and Symantec. All companies have global markets and their content is translated into many different languages. The number of languages varies from 20-100, depending on product types, regions and size of market.

The interviewees are professionals who participate in the decision making about content translation and localisation. Their roles in the companies vary: some participants described themselves as ‘Director of Localisation/Translation Services’, others as ‘Project Manager/Engineer’, or ‘Director of Translation Infrastructure’; and another as ‘Director of Research’.

2.2 Interview

The questionnaire consisted of two parts: 1) content profiling and 2) translation strategy (see below for specific questions).

The content profiling part aimed at identifying:

- What types of content the companies produce
- How they profile their content
- Into what languages the content is translated
- How the content types are translated
- The factors that decide whether content is translated, if at all.

We decided to use the eight meta-categories suggested by TAUS (see above) as the basis for the content profiles as we wanted to determine if the companies could fit their content into those categories and which content did not fit. (Note that at the time of the interviews, the TAUS profiling system had only eight categories whereas now it has ten, with ‘Legal’ and ‘Knowledge Base’ added more recently).

Our main objective was to identify common content profiles (if such existed) and the factors that drove decisions on how/whether content is translated.

The questions about content profiling were as follows:

1. Which of the following content types does your company produce? [TAUS categories]
2. Are there other content types that the company produces?
3. Do you produce content for internal purposes only? If so, what type of content and for what purposes?
4. For which of these communiction channels do you produce content?
    a) Business to Business
    b) Business to Consumer
    c) Consumer to Consumer
    d) Others

5. Into what languages are the content types normally translated?
6. What factors decide which target languages will be included for specific content types?
7. Which content types are translated by:
    a) HT only (i.e. human translation without Computer-Aided Translation - CAT)
    b) CAT only (i.e. TM and glossary tools)
    c) MT only
    d) MT + HPE (human post-editing)
    e) MT + APE (automatic post-editing)

8. What content types are never or rarely translated and why?
9. What factors are taken into consideration when deciding what will be translated (HT/MT) or left in the original?
10. Do you think the definitions of Utility, Time and Sentiment in the TAUS Framework are fit for your company's content? Does your company use other parameters? Are these concepts (TAUS's or your own company's) used in determining what should/should not be translated?

The second part of the interview, focussing on the relations between authoring and the translation strategy, aimed at identifying:

- How the translation process is performed

- How the authoring team is managed
- Whether there is collaboration between the authoring team and the translation team
- What evaluations are performed before and after the content types are published

By identifying the points above, we hoped to have an overview of the translation decision-making process so that we could design follow-up experiments on end-user reception of different types of translation (i.e. HT, MT only, MT+HPE).

The questions about translation decision strategies were as follows:

11. Does the company follow any specific style guidelines for translation? Do they differ across:

    a) content types
    b) users
    c) platforms

12. Is the quality of the translation always assessed before being published? How? (Which metrics?)

13. Is the concept of "Personas" used in authoring? If so, how does the translation process deal with the concept?

14. Does your company carry out any kind of end-user satisfaction evaluation, specifically for the content you produce? Or just for the products (without specific reference to content)? If not, why not? How is it done? Does it feed back into the style guide for translation?

15. How are authoring teams managed (in-house/freelance authors)? Is there any cooperation between the authoring team and the translation management team? If not why not?

16. What works well and what could be improved?

17. Is there a terminology management process in your company? Is there any collaboration between term management for English as the source language and for the translated content?

The next section provides an in-depth analysis of the answers collected.

3 Data Analysis

This section will report the results on content profiling and translation decisions collected from the interviews. Note that in this section, for confidentiality reasons, companies are given identifiers ‘A’, ‘B’ and so on, rather than their specific names.

3.1 Content Profiling

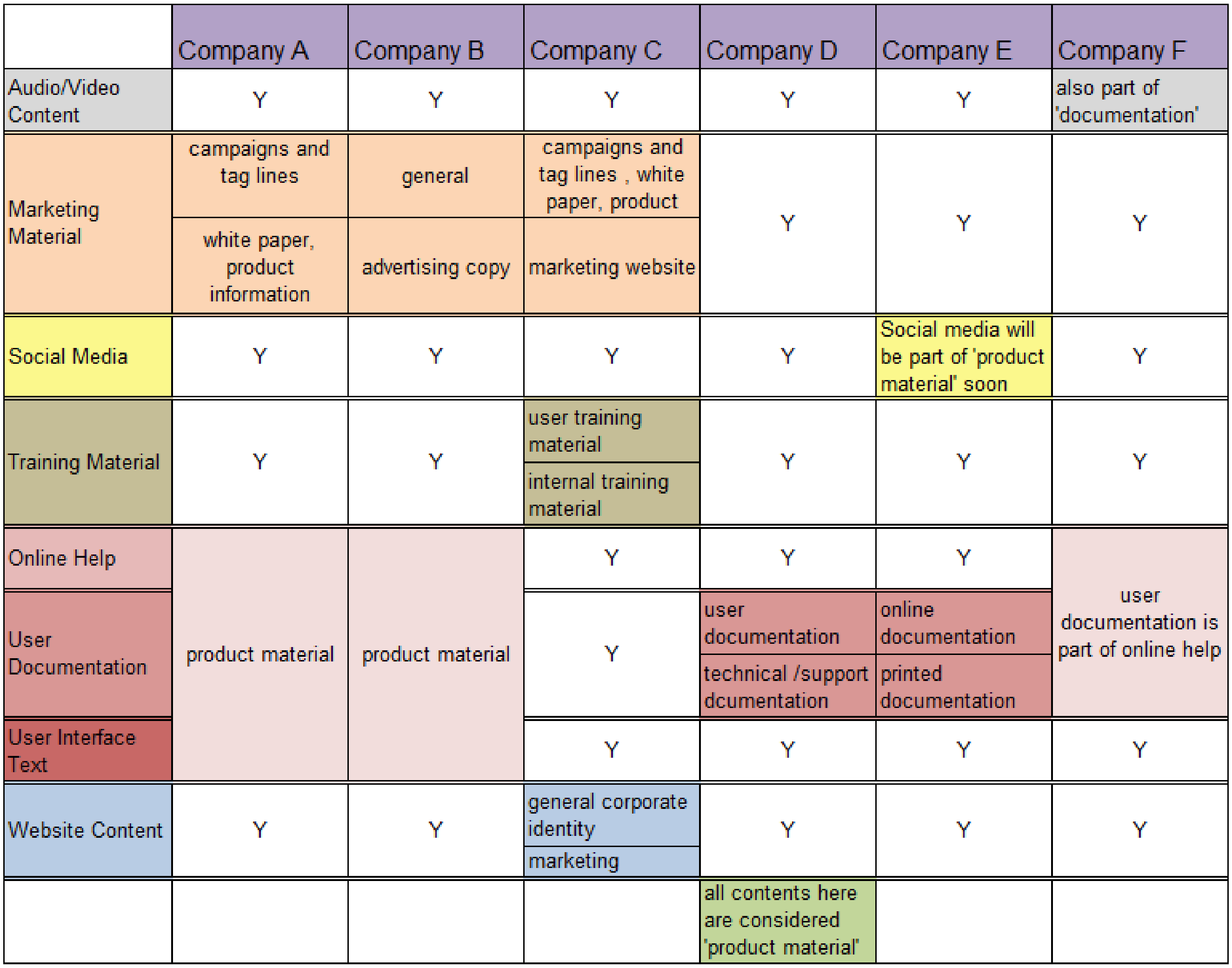
As mentioned previously, the first part of our questionnaire aimed at content profiling. The goal was to identify common content profiles by using TAUS meta categories as a starting point. All the participant companies confirmed that they produced all the content types listed in our first question. However, some companies categorise content differently, branching out into more detailed typologies, or combining content types into one single category.

It can be seen from the data in Table 1 that Online Help, User Documentation, User Interface, and Marketing Material are the content types that vary the most. Marketing Material is divided into different categories by Companies A, B and C. Note that Company C considers Marketing Material as one of the categories of Website Content:

> So the website would be the UI of the website and the actual marketing material on the websites [...]. Sometimes marketing would have a campaign website specifically for that [...] but our main website is just your corporate identity on the Internet.
>
> Company C

Company F considers User Documentation to be part of Online Help and Audio/Video content as part of ‘Documentation’. Companies D and F divide User Documentation into two

Table 1 - Content Profile per Company



	Company A	Company B	Company C	Company D	Company E	Company F
Audio/Video Content	Y	Y	Y	Y	Y	also part of 'documentation'
Marketing Material	campaigns and tag lines	general	campaigns and tag lines , white paper, product	Y	Y	Y
	white paper, product information	advertising copy	marketing website			
Social Media	Y	Y	Y	Y	Social media will be part of 'product material' soon	Y
Training Material	Y	Y	user training material	Y	Y	Y
			internal training material			
Online Help	product material	product material	Y	Y	Y	user documentation is part of online help
User Documentation			Y	user documentation	online documentation	
				technical /support dcumentation	printed documentation	
User Interface Text			Y	Y	Y	Y
Website Content	Y	Y	general corporate identity	Y	Y	Y
			marketing			
				all contents here are considered 'product material'		

different categories. Online Help, User Documentation and User Interface are grouped into **'product material'** by Companies A and B, while for Company D all of the eight content types would be the 'product material'.

Finally, Company E considers that Social Media will be part of 'product material' soon; although it is not very clear what other content types would be included as such.

All the participant companies mentioned content types which were not listed in the eight starting categories. They are:

- Employee Engagement Survey
- Internal Announcements
- Support Documentation
- Online Knowledge Base
- Legal Texts
- Surveys (customers and end-users)
- User Generated and Industry Generated Content
- Sales Training Material
- Internal Sales Tools Texts
- Internal Training Material
- Metadata
- Templates
- Technical Developer Documentation

Some companies have a more detailed classification than our baseline. For example, for Training, some companies consider the ‘Training Material’ in our classification for end-users (business and costumers) only. Therefore, Sales Training Material is for the sales team only and Internal Training Materials are for internal employees only. Note that Legal Texts and Online Knowledge Base fit the two categories added subsequently to TAUS profiling system.

When asked if they produced content for business and consumers (Question 4 in Section 2), all participants answered that they would produce content for all three options. Only one of the participants said they would translate Consumer-to-Consumer content types (e.g. online forum content), while the remaining either did not know or did not translate such content. Business-to-Business and Business-to-Consumer content is always translated.

We asked the participants whether or not they agreed with the TAUS dynamic quality evaluation parameters of utility, time and sentiment (Question 10) and whether they believed those parameters could be applied to their content and subsequent translation decisions. All companies reported that **Utility** is a very important parameter and could be applied in profiling their content. Companies A and B reported that **Time** was not a factor since ‘everything must be fast’, while Company D reported time as ‘becoming very important for the new world’ where new features of software are released weekly and need to be translated at the same time of their release. Company C said that **Time** comes together with **Utility**.

Regarding **Sentiment**, Company A said it is directly connected to brand image while Company B said **Sentiment** would fit only for Social Media content. Companies E and F said all three parameters would fit their content.

We also asked the participants whether they had other parameters in use to profile their content on how, where and whether or not to translate them. They were:

- Cost
- Quality
- Region

Quality and Cost were the most frequently mentioned parameters. Quality was mentioned by four companies while Cost was mentioned by five of them. Region refers to the size of the market in different countries.

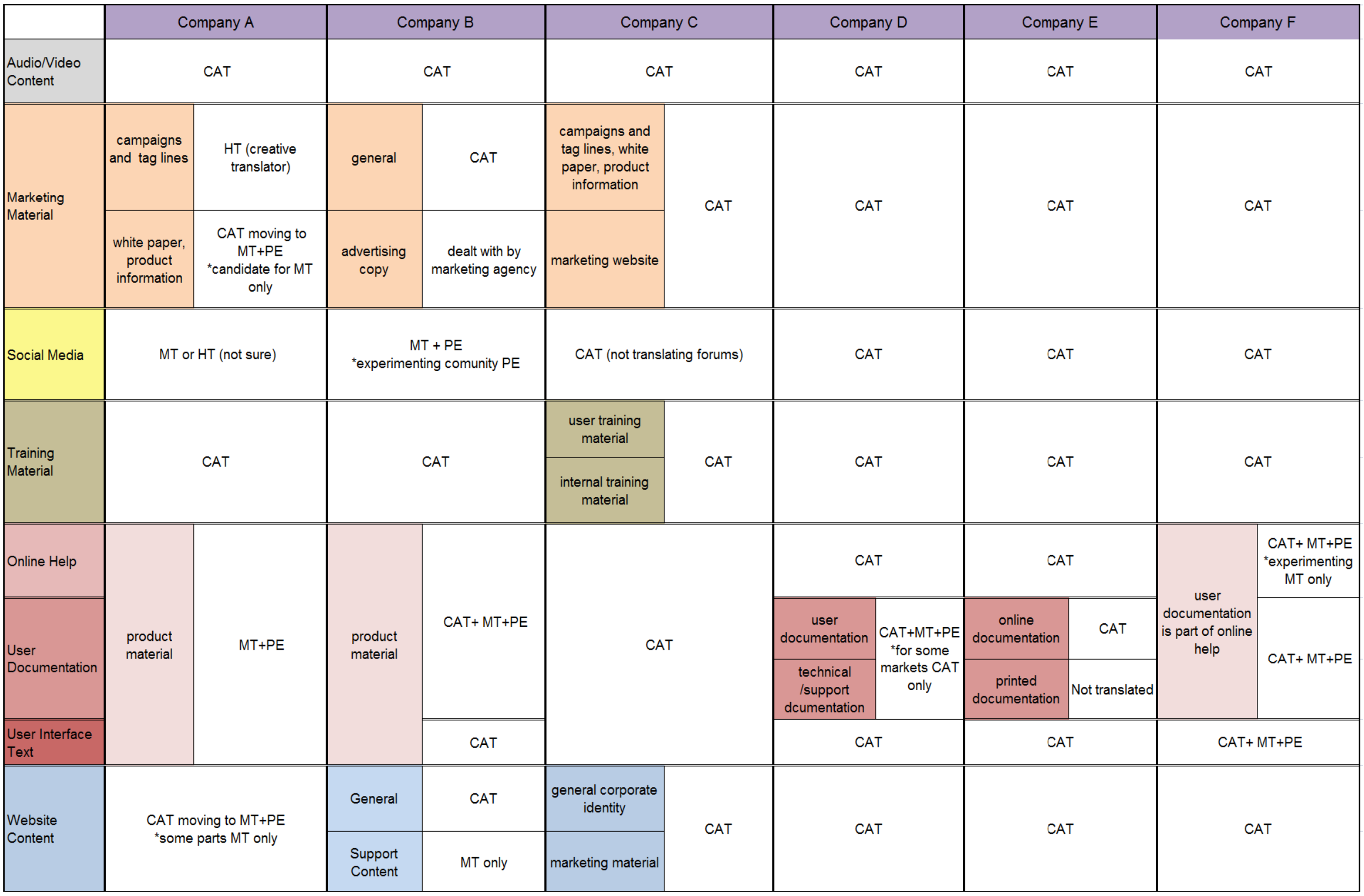
*3.1.1 Content Translation***.**

One of our goals was to identify how the companies translated their content. During the interviews, we gave our participants a list of translation methods and asked them to name which content type was translated by CAT (computer-aided translation), HT (human translation), MT (machine translation), MT+PE (machine translation plus some kind of post-editing), or left in the original. Note that CAT refers to the use of TM (and glossaries) only; HT refers to human translation only, with no use of CAT or MT; MT refers to the use of raw machine translation output only, with no kind of post-editing.

Table 2 shows responses in detail and Figure 1 summarizes the results comparing each content type and their translation types. As can be seen from Table 2 and Figure 1, the majority of content types are currently translated using CAT tools. This is unsurprising given that CAT tools have been very common in the localisation market for decades.

Company A, which uses 'HT' for marketing campaigns and tag lines, is the only company to use HT solely.

Table 2 - How Each of the Content Types Are Translated per Company



	Company A		Company B		Company C		Company D		Company E		Company F	
Audio/Video Content	CAT		CAT		CAT		CAT		CAT		CAT	
Marketing Material	campaigns and tag lines	HT (creative translator)	general	CAT	campaigns and tag lines, white paper, product information	CAT	CAT		CAT		CAT	
	white paper, product information	CAT moving to MT+PE *candidate for MT only	advertising copy	dealt with by marketing agency	marketing website							
Social Media	MT or HT (not sure)		MT + PE *experimenting comunity PE		CAT (not translating forums)		CAT		CAT		CAT	
Training Material	CAT		CAT		user training material	CAT	CAT		CAT		CAT	
					internal training material							
Online Help	product material	MT+PE	product material	CAT+ MT+PE	CAT		CAT		CAT		user documentation is part of online help	CAT+ MT+PE *experimenting MT only
User Documentation							user documentation	CAT+MT+PE *for some markets CAT only	online documentation	CAT		CAT+ MT+PE
							technical /support dcumentation		printed documentation	Not translated		
User Interface Text				CAT			CAT		CAT		CAT+ MT+PE	
Website Content	CAT moving to MT+PE *some parts MT only		General	CAT	general corporate identity	CAT	CAT		CAT		CAT	
			Support Content	MT only	marketing material							

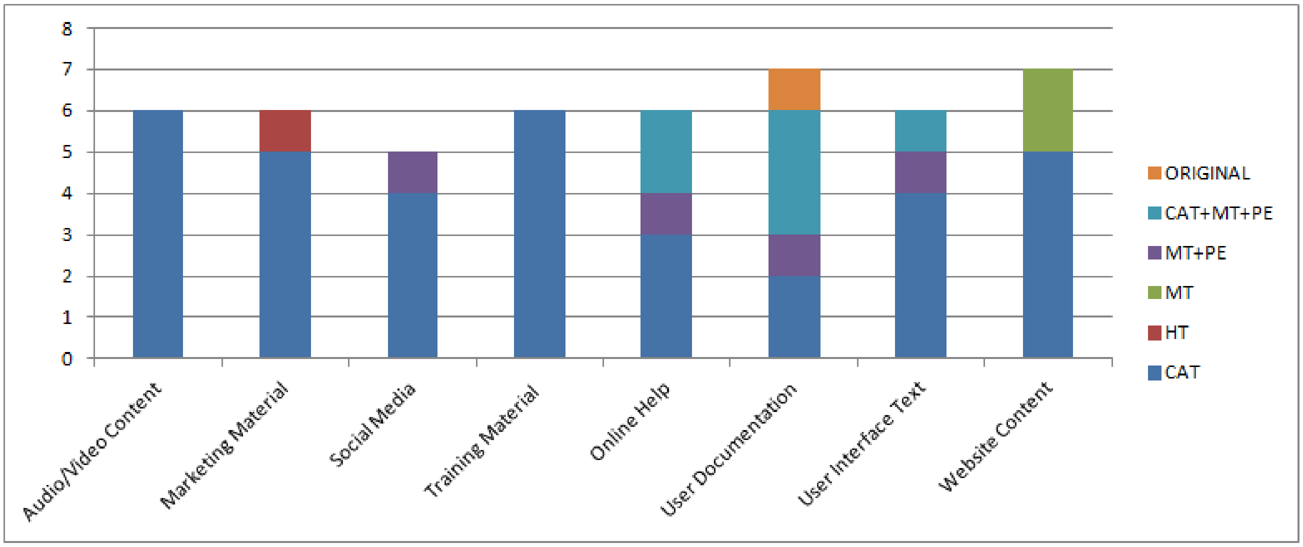


Figure 1 – Content Types vs. Translation Mode

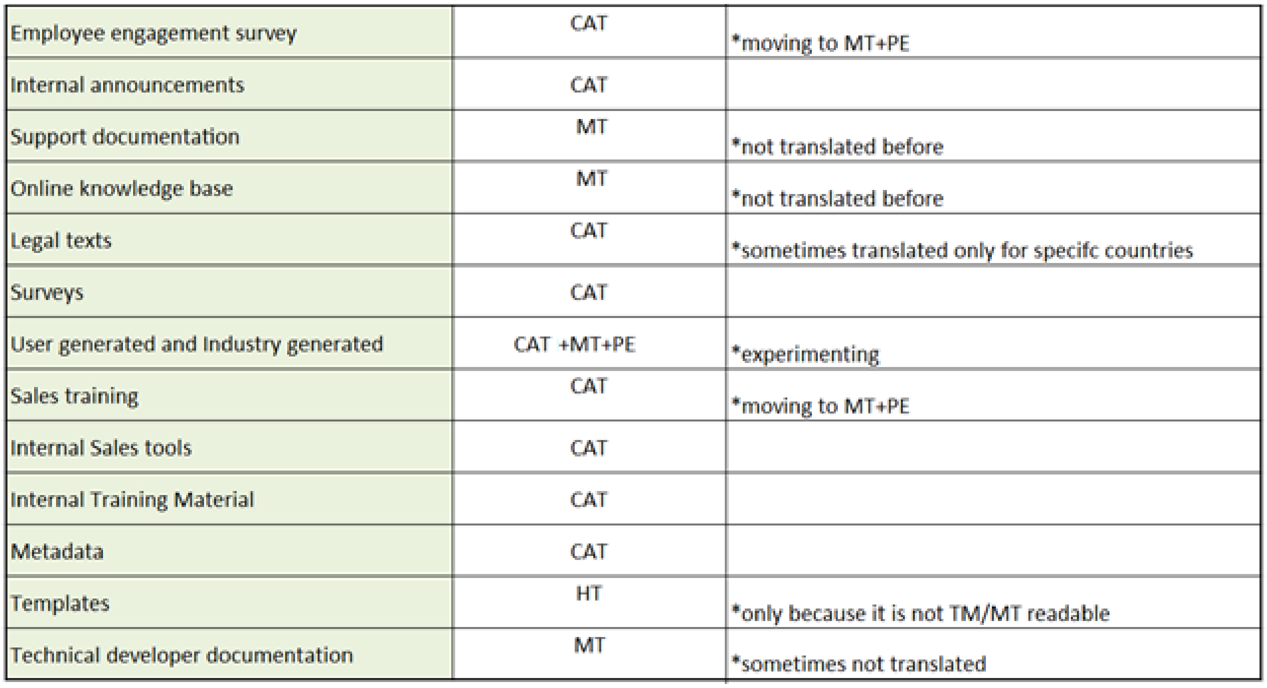
Raw MT (MT only) is used by Companies A, B and F for Website Content and Online Help. However, it is interesting to note that the amount of raw MT being used to translate content is still very low. 'MT only' is used for the 'support content' or 'some parts only' of the website content and it is still in an experimental phase for Online Help (Table 2). Note that Company A, dividing marketing material into 2 categories, reports that white paper and product information of marketing material "are very good candidates for MT".

A common practice for translating some of the content types is to use CAT+MT+PE. The content is first translated with the TM database and the sentences that are not matched will be then machine translated and post-edited.

Regarding the content types that were not listed in our questionnaire (see section 3.1), the results were similar to the ones in Table 2. Table 3 provides a summary.

However, it is interesting to note that for these content types, there is more ongoing experimentation. Some content types are translated by CAT tools but it is foreseen that they will be translated by MT+PE in the future. Also, some content types which were not previously translated are now being translated by MT only (e.g. knowledge base, technical developer documentation)

Table 3 - How not-listed content types are translated



Employee engagement survey	CAT	*moving to MT+PE
Internal announcements	CAT	
Support documentation	MT	*not translated before
Online knowledge base	MT	*not translated before
Legal texts	CAT	*sometimes translated only for specifc countries
Surveys	CAT	
User generated and Industry generated	CAT +MT+PE	*experimenting
Sales training	CAT	*moving to MT+PE
Internal Sales tools	CAT	
Internal Training Material	CAT	
Metadata	CAT	
Templates	HT	*only because it is not TM/MT readable
Technical developer documentation	MT	*sometimes not translated

Note that the content types that are translated with MT only were either 'not translated before' or 'sometimes not translated'. Those content types involve more technical content than the others.

According to the participants, the decisions on how/whether content types will be translated depend on a series of factors. The keywords used by participants in discussing these factors were:

- Brand Image
- Business Case
- Cost
- Profit

- Region
- Return on Investment
- Revenue
- Size of the Market
- Strategic
- User Behavior/Audience
- Volume
- Effort
- Rating

Cost and ‘Strategic’ are the most cited factors (4 times), followed by Region and Return on Investment (2 times). Strategic refers to “whether there's growth in the market” and, according to the participants, it is a political decision.

Volume refers to the amount of text to be translated and Rating to the average rating of content (on websites pages) to decide if it will be translated or not. Effort refers to “how easy it is going to be to translate and how much work we're going to have to do to get a glossary done”. Several participants also commented on how defining the audience could help decide how to set the quality level expected by the user:

The more we know about where that content is going, to the audience, the more we can set a quality level in the metadata and we can take more risk in the work flow.

> It's the content type, hence the audience type. Sometimes it is by market. For example, in general, you find that even though the overall [translation] quality may be the

> same or similar, [Language A] users tend to be more accepting of MT than [Language B] users. And [Language C] users are actually more accepting of MT than [Language B] users, so that's also a factor
> *the languages were removed to preserve the company's identity

According to Rockley and Cooper (2012, 67), companies "need to determine how well your current content is meeting your customers' needs and identify any gaps in the content".

Setting strategic guidelines and assessing the content (source and translated) seem to be an important step to determining customer's needs. The next section will report on the answers about guidelines and assessment for both source and translated content.

3.2 Translation Strategy

The second part of the interview aimed at identifying the translation decision-making process inside the enterprise. In this report, we focused on guidelines and evaluation for both translation and authoring.

3.2.1 Translation.

We asked our participants some questions about their translation process:

- **Guidelines**

When asked if the company produced any specific guidelines for translation, all participants confirmed they had some kind of guidelines for translation. Some of the participating companies have general guidelines with general rules and others have very specific ones, for different markets/region or content.

- **Translation evaluation**

The majority of respondents reported that they had some kind of evaluation process for translated content. Companies A, B, D and F said they would have a linguistic review. For some, it would be a spot check review, for others all translated content would be revised, depending on the product and the content type. Company C reported they use the LISA QA model in some samples of the content.[1] Only one company (E) said they do not do translation evaluation.

- **End-user evaluation**

A small number of those interviewed reported that they had some kind of end-user evaluation. This evaluation commonly focuses on the product and website (usability). Online surveys may include one or two questions about translation but mostly, they are about the 'content'. However, one of our participants claims that online surveys may give feedback for translation as well:

> - Is there any question about the translation, the language itself? [in the online survey]
> - I suppose inherently because lots of the questions are about what your experience about the content was. But whether actually it uses the word translation I'm not sure but if it is asking you in French about your experience of the content, the content is in French... feedback is going to tell you if there is any problem with the translation

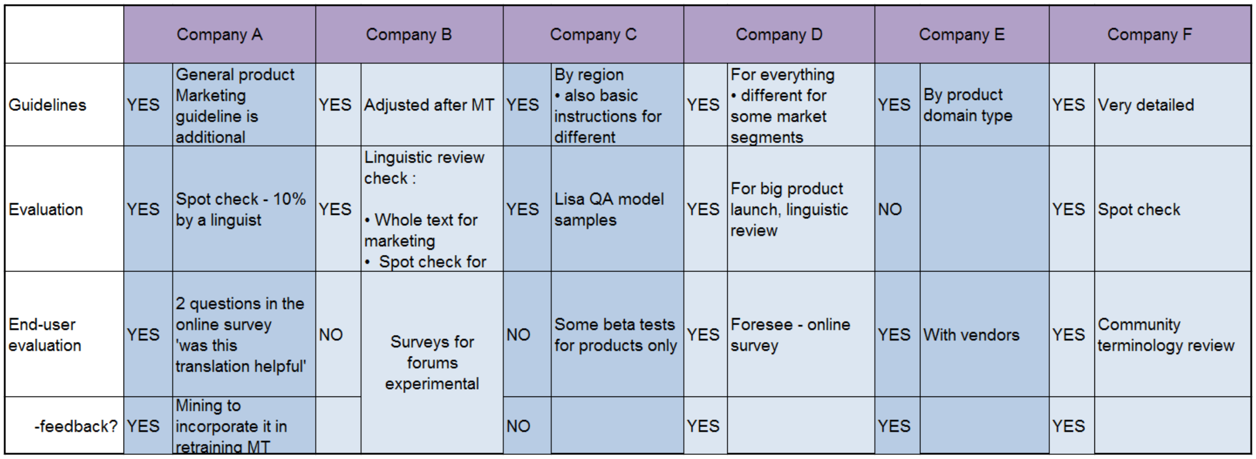
It is interesting to note that Company A, which has two questions about the translation in their online survey, is mining the feedback obtained within the survey to retrain their MT system.

[1] LISA was the Localisation Industry Standards Association, which is now defunct, but whose translation QA model was largely adopted by localisation companies.

Company F uses community terminology review and Company E has some evaluation directly with their vendors.

Table 4 summarises the answers about guidelines, and evaluation for translation.

Table 4 – Translation Guidelines and Assessment Practices per Company



	Company A		Company B		Company C		Company D		Company E		Company F	
Guidelines	YES	General product Marketing guideline is additional	YES	Adjusted after MT	YES	By region • also basic instructions for different	YES	For everything • different for some market segments	YES	By product domain type	YES	Very detailed
Evaluation	YES	Spot check - 10% by a linguist	YES	Linguistic review check : • Whole text for marketing • Spot check for	YES	Lisa QA model samples	YES	For big product launch, linguistic review	NO		YES	Spot check
End-user evaluation	YES	2 questions in the online survey 'was this translation helpful'	NO	Surveys for forums experimental	NO	Some beta tests for products only	YES	Foresee - online survey	YES	With vendors	YES	Community terminology review
-feedback?	YES	Mining to incorporate it in retraining MT			NO		YES		YES		YES	

3.2.2 Authoring

We asked our participants some questions about authoring as we wanted to understand the content creation process before it is handed to the translation team. The majority of the participants responded that they had in-house authors. A few of them have a small percentage of outsourced authors.

- **Guidelines**

When asked if the company had any guidelines for authoring, all participants confirmed that they had some guidelines for source content. One interesting observation is that some participants have their source content guidelines tuned for machine translation and also have their guidelines adapted every now and then:

> We adjusted the style guides when we brought in machine translation and it was a really healthy thing that we did at the time

> we're changing the source writing so that it's simpler

However, the participants claim that writing guidelines for authors is a very hard task. Because of the existing gap between groups inside the same company (the aforementioned ‘silos’), it makes it hard to set some rules:

> writers tend to feel like creative people and they don't really want to be told how to do stuff

> it is always a compromise because they [the authoring team] want to specialise and we want to standardise.

- **Cooperation with Translation Team vs. Silos**

Even when there is cooperation, it may not be from all the authoring groups since the companies may have different groups for each product/domain.

> They [the authoring team] are decentralized, so they are in different product groups

Even though all the respondents confirmed that there is some kind of cooperation between the translation and authoring teams, they frequently report that the cooperation is between a small number of authoring teams only and that they are actively ‘trying to bridge the gap’ between both worlds.

- **Evaluation**

Regarding source evaluation, because most of the participants did not mention it during the interview, we decided to ask some follow-up questions[2]:

a) How do you identify bad quality source text?

b) Is the source content published before translation?

c) Is the feedback from translators the factor that decides if the content is bad?

d) What happens to bad quality source? Is it sent back to the authoring team?

e) Are translators expected to correct the source while translating?

Table 5 summarizes the answers to these questions.

Table 5 - Source Evaluation Practices per Company

	Company A	Company B	Company C	Company D	Company E	Company F
How do you identify the bad quality of the source?	X	Acrolinx Trained to their style guide. Bad Acrolinx score will predict a high PE/Translation effort	Not done Translators point out issues (queries) and the queries are tracked for later analysis	Copy-editing	Automated validation checks Their own tools	Automated validation checks Their own tools
Is the source content published before translation?	X	Published Simultaneously	No	Big launches - simultaneously lower priority - sometimes english may be first	Published Simultaneously	Big launches - simultaneously
Is the feedback from translators the factor that decided if the content is bad?	X	One of the factors Preparation phase gets most of the issues	Yes	One of the factors, but minor event. Copy-editing gets most of the issues	One of the factors Preparation phase gets most of the issues	One of the factors Preparation phase gets most of the issues
What happens to bad quality sources? Do you send it back to the authoring team?	X	Source is sometimes sent back Feedback is sent while translating	Source is sometimes sent back (query system) If there is time or if the error is misleading to the user	Source is sometimes sent back but just in case of severe problems (very rare). Copy-editors are supposed to correct.	A file cannot enter the translation process unless it is passed as valid by these tools.	Source is sent back if it does not pass the automated validation checks
Are translators expected to correct the source while translating ?	X	Translators should try to address the issue without changing the source This misallígns the TM matches in the future	Translators correct the translation but don't change the source This misallígns the TM matches in the future	Translators don't correct bad source	Translators don't correct the source. This would break one of the fundamentals of source control – translation management.	Translators don't correct the source. Translators my handle errors in the source with the translation and feedback is sent

[2] All Companies but Company A had time to answer the questions.

Regarding how the Companies identify bad quality source, Companies B, E and F use automated validation checks; copy-editing is used by one Company (D) and source evaluation is not done before sending to translation by Company C.

Companies B and E said they publish the content simultaneously (question b); Companies D and F said they publish simultaneously if it is a big product launch otherwise the English is published first, and one (Company C) said source is always published first.

When asked if the feedback from translators is the factor that mostly decides the quality of the source only Company C answered ‘yes’. The other Companies said translators’ feedback is only one of the factors, as the preparation phase (automated or copy-editing) should identify most of the issues.

Regarding sending bad quality source content back to the authoring team, most of the Companies said they ‘sometimes send source back’ and the reasons for that vary greatly. Company B stated that the translation of the source starts a little after the source creation starts; therefore, creation and translation happen almost simultaneously. Feedback from the translators is sent while translating. Company C said bad quality source is sent back only when there is enough time or, even when time is an issue, if the source is misleading the user, if has to be sent back. Company D stated that source is sent back only in case of severe problems, however, this is a very rare event as the copy-editors should correct those issues. Companies E and F said the source does not enter the translation process unless it is validated by the automated checks.

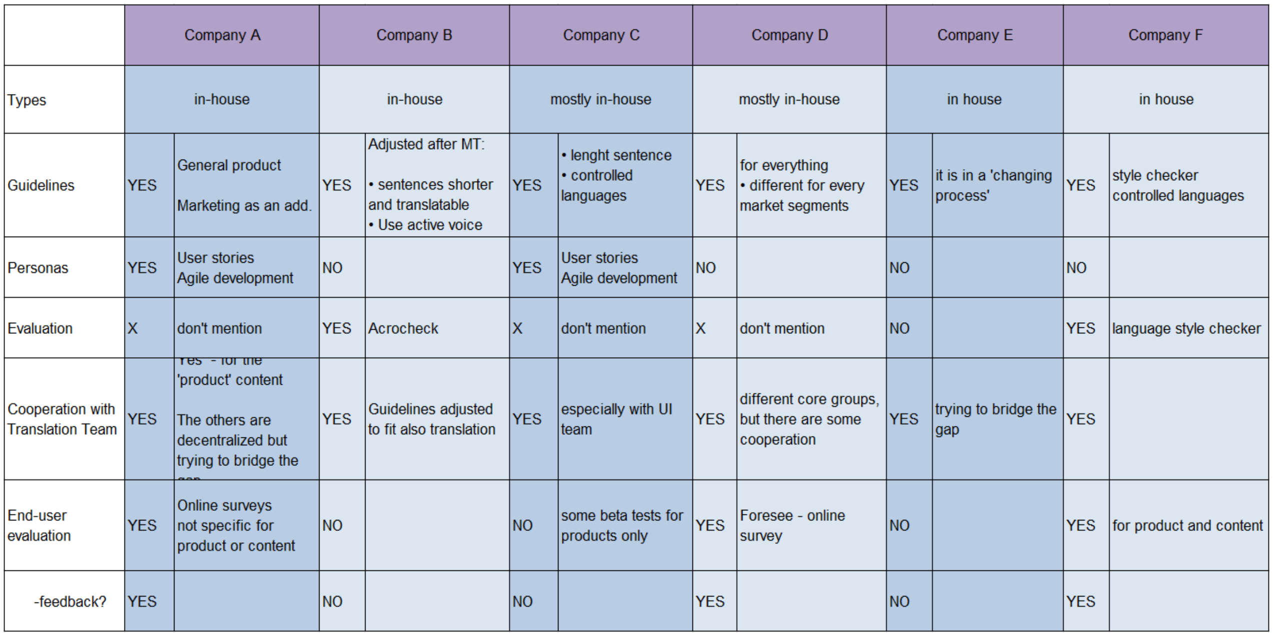
Finally, when asked if translators are expected to correct the source, all companies said the translators should not correct the source, but they should handle any issues that may make it to the translation process. Again, the reasons for that vary hugely. Companies B and C declared that translators cannot change the source as “this misaligns the TM matches in the future”. Company D stated that translators cannot correct the source since it is the copy-editors’ job to do so. Companies E said translators are not supposed to correct source since this “would break one of the fundamentals of source control – translation management”. And Company F stated that translators do not correct source and if any errors make it into the translation process, feedback is sent to the authoring team.

- **End-user evaluation**

End-user evaluation is also another point that seems to be under-deployed. Only one company said they would do end-user evaluation both for product and content.

Table 6 summarises the answers about guidelines, evaluation and cooperation with the translation team.[3]

Table 6 - Authoring Practices per Company



	Company A		Company B		Company C		Company D		Company E		Company F	
Types	in-house		in-house		mostly in-house		mostly in-house		in house		in house	
Guidelines	YES	General product Marketing as an add.	YES	Adjusted after MT: • sentences shorter and translatable • Use active voice	YES	• lenght sentence • controlled languages	YES	for everything • different for every market segments	YES	it is in a 'changing process'	YES	style checker controlled languages
Personas	YES	User stories Agile development	NO		YES	User stories Agile development	NO		NO		NO	
Evaluation	X	don't mention	YES	Acrocheck	X	don't mention	X	don't mention	NO		YES	language style checker
Cooperation with Translation Team	YES	Yes - for the 'product' content The others are decentralized but trying to bridge the gap	YES	Guidelines adjusted to fit also translation	YES	especially with UI team	YES	different core groups, but there are some cooperation	YES	trying to bridge the gap	YES	
End-user evaluation	YES	Online surveys not specific for product or content	NO		NO	some beta tests for products only	YES	Foresee - online survey	NO		YES	for product and content
-feedback?	YES		NO		NO		YES		NO		YES	

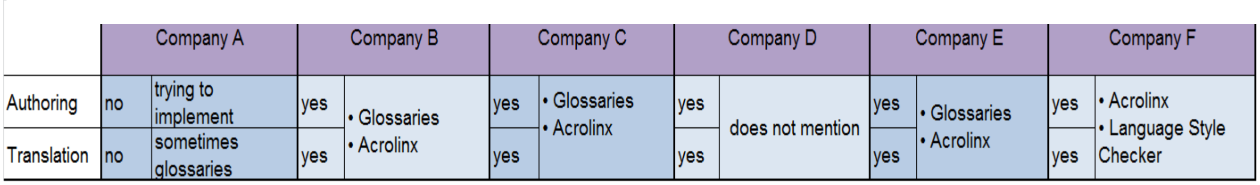
[3] Note that for Source Evaluation, Table 6 displays the data collected before we decided to send the follow-up questions to the participants.

3.2.3 Terminology Management

Regarding Terminology Management, almost all participants reported they had a Terminology Management process for the source and for the translated content. An interesting fact is that all of the companies seem to have implemented the terminology process a short time ago and are now adjusting it. Even the companies that said they do not have a terminology management process confirmed they are currently trying to implement one.

Acrolinx and Glossaries seem to be the most common method used by the companies for managing terminology. (Note: Acrolinx is described as 'content optimization software' that increases the readability and translatability of content (www.acrolinx.com)). Table 7 summarises the results by company.

Table 7- Terminology Management per Company



	Company A		Company B		Company C		Company D		Company E		Company F	
Authoring	no	trying to implement	yes	• Glossaries • Acrolinx	yes	• Glossaries • Acrolinx	yes	does not mention	yes	• Glossaries • Acrolinx	yes	• Acrolinx • Language Style Checker
Translation	no	sometimes glossaries	yes		yes		yes		yes		yes	

4 Conclusions and Future Work

The purpose of the current study was to determine how companies in the IT localisation sector profile their content and how translation decisions are taken. We interviewed a small sample of key decision-makers in the localisation sector, but these decision-makers represent large multinationals who translate billions of words per year into many languages.

We have shown that regarding content profiles (Section 3) although some companies produce the same types of content, and although they say that their content fits with the TAUS list of content types, there is no consistency in how they profile them. Content types have different categories and companies also have different parameters for profiling them such as cost, quality and region.

Regarding content translation, we have shown that HT only is almost never used. Also, most content types are currently translated using CAT tools, and so we can say that this is the 'norm'. It is evident, however, that all the companies are either starting to experiment with MT (and/or CAT+MT+PE) or are already using it for their content. It also seems that the content types where MT is being used solely are predominantly technical. The decision on whether to use MT appears to be guided by the following: i. when the user expectation of quality is not very high, e.g., technical documentation is expected to have end-users with more tolerance for MT errors; ii. a content type that was not translated before due to cost or effort may be a good candidate for MT only.

All companies seem to translate what some call 'product material' – which is Online Help, User Documentation and User Interface – by default. When not translated by default, there are a number of determining factors that decide whether the content will be translated, how it is going to be translated, and into what languages, such as:

- brand image (relating to Social Media, the product box, or even all the content types)
- business case (cost, profit, return on investment, revenue)
- geographical factors (strategy, region, size of market)
- user-centric factors (user behavior, audience, rating)
- product-specific factors (volume, effort)

The companies also share the fact that they have guidelines for translation and authoring but they vary considerably regarding evaluation and, in particular, end-user evaluation. The latter has proven to be almost non-existent.

It is evident that there are no set guidelines for content profiling or for the translation decision making process. All decisions depend on a number of factors that may or may not be replicated with each newly launched product or new commercial region. To be able to map all the routes to the decisions taken about whether content is translated and via which process we would have to interview people from other parts of the enterprise, since many participants confirmed that most decisions were based on business factors.

Finally, as mentioned previously, all companies seem to be experimenting and 'trying to use more MT' and to implement it for more content types. It would be interesting to interview the participant companies in a year to observe how those changes (if any implemented) have taken place and how they have affected content profiling and the translation strategy.

5 References

Boiko, Boiko. 2005. *Content Management Bible*. John Wiley & Sons.

DePalma, A. Donald, and Benjamin B. Sargent. 2013. “Transformative Translation.” *Report.* Common Sense Advisory.

DePalma, A. Donald, Hegde, Vijayalaxmi, Pielmeier, Hélène, and Robert G. Stewart. 2013. “The Language Services Market: 2013.” *Report*. Common Sense Advisory.

Moorkens, Joss. 2012. “Measuring Consistency in Translation Memories: A Mixed-Methods Case Study.” *PhD dissertation*. Dublin City University.

O’Brien, Sharon, Choudhury, Rahzeb, Van der Meer, Jaap, and Nora Aranberri Monasterio. 2011. “Dynamic Quality Evaluation Framework.” *TAUS Labs Report*. The Translation Automation User Society - TAUS.

Rockley, Ann, and Charles Cooper. 2012. *Managing Enterprise Content: A Unified Content Strategy*. New Riders.

Smith, Heather A., and James D. McKeen. 2003. “Developments in Practice VIII: Enterprise Content Management.” *Communications of the Association for Information Systems 11*, article 33.