The efficacy of terminology-extraction systems for the translation of documentaries

RESEARCH ARTICLE

The efficacy of terminology-extraction systems for the translation of documentaries

Sabien Hanoulle[a]*, Véronique Hoste[b] and Aline Remael[c]

[a] Faculty of Arts, Department of Applied Linguistics, Translators & Interpreters, University of Antwerp, Belgium; [b] Faculty of Arts and Philosophy, Department of Applied Language Studies, Ghent University, Belgium; [c] Faculty of Arts, Department of Applied Linguistics, Translators & Interpreters, University of Antwerp, Belgium.

To cite this article: Sabien Hanoulle, Véronique Hoste & Aline Remael (2015): The efficacy of terminology-extraction systems for the translation of documentaries, Perspectives: Studies in Translatology, DOI: 10.1080/0907676X.2015.1010549
To link to this article: http://dx.doi.org/10.1080/0907676X.2015.1010549

Abstract

This article investigates whether the integration of a domain-specific, bilingual glossary supports audiovisual translators of documentaries in terms of translation process time and terminological errors. After a short review of issues typical of documentary translation and a discussion of the use of translation-memory software in general, the reference corpora are described. Next, a manually labelled glossary is created and its constitution is explained with special emphasis on the criteria used to qualify what is a "term" or not. This glossary is then used as a gold standard to calculate the rate of agreement with the glossary of three automatic terminology-extraction systems. Finally, experiments with master students demonstrate how both glossaries (the gold standard and one automatically extracted glossary) reduce their process time significantly but not the number of terminological errors. The article concludes with a discussion of the data analysis and by presenting the next step in this research i.e. experiments with professional translators and further challenges such as a comparison between the glossaries.

Keywords*:* documentaries; audiovisual translation; off-screen dubbing; terminology-extraction systems; translation-memory software

Introduction

Over the past fifteen years scholarly studies have shown a growing interest in the translation of documentaries, highlighting the versatility of this audiovisual translation mode. Translating documentaries requires an all-round knowledge of text types and functions, language registers, domain-specific terminology and translation techniques such as voice-over, dubbing, off-screen dubbing and subtitling. As Franco (2000) demonstrates, it is a 'specific practice' (p. 234), and Espasa (2004) argues

that it is also specifically audiovisual in the challenges it poses for translation and research, one of which is a varied target audience in terms of age and cultural background, which also poses challenges for terminology translation.

Matamala (2009) agrees that terminological problems occupy a prominent position in the translation process as there can be more than one equivalent term in the target language or none whatsoever. Conducting corpus-based research into this specific field, she has identified terminological challenges including 'identifying terms, understanding terms, finding the right equivalent, dealing with the absence or the inability to find an adequate equivalent, dealing with denominative variation, choosing between in vivo and in vitro terminology and avoiding wrong transcriptions' (Matamala, 2010, p. 269). She suggests, with respect to terminological neologisms when no adequate equivalent is found, that 'it would be highly interesting [...] to find a mechanism to avoid duplication of efforts' (2010, p. 269). As a matter of fact, these mechanisms exist. Translation-memory (TM) software stores and matches source and target text sentences or sub-sentential segments for future reuse, sometimes combined with terminology extraction and management. In an ethnographic study in Canadian translation services, Le Blanc (2013) describes that in those services, TM software is used for nearly all types of texts (general, administrative, technical and specialised). However, in an overview of representative empirical TM studies, Christensen and Schjoldager (2010) demonstrate that the main text types used are technical, scientific, financial and legal texts. The strong correlation between these text types and the use of TM systems, analysed by Lagoudaki (2010), is due to a high degree of repetition and a large amount of terminology. An example of the importance of terminology management can be found in Coombs' (2014) case study of biotechnological patent translation. Thanks to a centralized translation service provider, the company not only reduced translation costs but also translation errors and it could thus improve quality.

Documentaries, however, contain a mixture of general utterances and domain-specific terminology. Therefore, this paper aims to investigate whether translation memory tools, and more specifically terminology-extraction software, support audiovisual translators of documentaries as well.

Research questions and hypotheses

The main research question is: 'Does the integration of bilingual glossaries into the translation process reduce translators' workload and/or the number of terminological errors they make?' This general question can be operationalized by subdividing it into three more concrete questions:

(1) Does an exhaustive, manually labelled bilingual glossary (called gold standard in this article) help to translate documentary texts as far as process time and number of terminological errors are concerned?
(2) Is the terminology used in documentaries specific enough to be detected by automatic terminology-extraction systems and can these systems generate a bilingual glossary that comes close to the gold standard? How accurate is this glossary compared to the gold standard?
(3) Does an automatically extracted bilingual glossary help to translate documentary texts as far as process time and number of terminological errors are concerned?

As researching this terminology for translation slows down the translation process, our hypothesis was that the integration of a bilingual, domain-specific glossary would both speed up this process and make it less error-prone. Moreover, we hypothesized that natural science documentaries may contain a mixture of standard language and technical terms, but also contain domain-specific terminology, and may therefore be eligible for accurate, automatic term detection.

Organisation of the research and reference corpora

This article describes the two main stages of the research conducted so far: the preparatory stage (in which bilingual glossaries were drawn up) and the experiment proper (the translation test). In order to verify whether the main corpus of this study contains domain-specific terminology, the terms of a sub-corpus of ten texts were labelled manually and ranked based on their degree of relatedness to the context. Next, we created a bilingual glossary called gold standard, labelling the terms of a sub-corpus of three texts manually. This gold standard was then compared to the automatically extracted glossaries of three terminology-extraction systems. The experimental part consisted of a translation test conducted with master students in translation, working without and with the bilingual glossaries. The article concludes with a discussion of the results and proposals for further research.

The main corpus used in this study was limited, in terms of genre, to natural sciences, as these episodes presumably contained enough recurring, domain-specific terminology to be considered for accurate, automatic term detection. Out of the whole corpus of natural science documentaries, the three most frequent subjects were selected: wildlife, earth & space and human body. In terms of audiovisual modes, only off-screen dubbing[1] has been considered. Every audiovisual mode (subtitling, dubbing, voice-over, off-screen dubbing and audio description) has to deal with specific constraints which might influence the translational choices, hence the need to focus on one mode. In subtitle

[1] Term proposed by Franco, Matamala & Orero (2010) which indicates the translation of the commentary voice heard off-screen.

countries the most frequent translation mode for documentaries is off-screen dubbing often combined with subtitles. The language combination of the corpus is English - Dutch, the major language pair on Flemish television and the target culture is Flanders, as the material was made available by the VRT, the Flemish broadcaster. The main corpus consists of the original English scripts of 171 episodes and their translations, all broadcast between 2005 and 2011. Sub-corpus one consists of ten representative episodes and was used to check whether the main corpus contains domain-specific terminology.

Sub-corpus two consists of three representative episodes that served to create a gold standard*:* *The earth machine - Land, Madagascar – Island of marvels* and *The secret world of pain*. Excerpts from these episodes also served as a source text for the experiment.

The preparatory stage

Manual labelling of terms

In order to check whether the main corpus of this study contains domain-specific terminology, all source text and target text sentences of sub-corpus one were aligned by means of a translation-memory software, Similis® (http://similis.org/linguaetmachina.www/index.php), and the alignment was verified manually. Next, one annotator labelled and ranked all the terms into three sub-categories, based on the degree of relatedness to the domain-specific context where sub-category one stands for very strongly related to the context, two strongly related and three not so strongly related but still related. The ten texts contained an average of 2.25% term types for sub-category one, 1.80% for sub-category two and 2.11% for sub-category three, which might seem small numbers. However, these terms can be very specialized, for example ‘giraffe necked weevil’ *(Madagascar)* or ‘mantle plume’ *(The earth machine),* so it is fair to say that a bilingual glossary constitutes an additional resource of specialized information that might give support to translators.

To investigate the efficiency of a glossary, an exhaustive and accurate bilingual glossary, called gold standard (GS), was created. In order to do this, the source and target texts of sub-corpus two were aligned and three annotators labelled all the domain-specific terms manually. The final GS was constituted by assembling all the terms identified by the three annotators.

However, manual labelling of terms can only be carried out on the basis of a clear-cut definition of what a ‘domain-specific term’ is. The following definitions highlight two different theoretical perspectives, both relevant to this study. According to Wright (1997), '[T]erms [...] are the words that are assigned to concepts used in the special languages that occur in *subject-field or domain-related texts*' (p.13, my italics). Bowker adds a linguistic angle: 'Terms consist of *single-word or multi-word units*

that represent discrete conceptual entities, properties, activities or relations in a particular domain.' (Bowker, as cited in Macken, 2010, p. 104, my italics). A common denominator in these definitions is 'concept'. Macken, Lefever and Hoste (2013) provide three criteria to qualify a term in a domain-specific text:

1. Termhood refers to 'the degree to which a linguistic unit is related to a domain-specific context' (Kageura & Umino, 1996, p. 260-261) i.e. 'each entry in the extracted lexicon should refer to an object or action that is relevant for the domain'. This means that the frequency of the extracted linguistic unit versus other domains or versus general speech will be higher. For instance the word 'bird' in a documentary about the flora and fauna in *Madagascar* is considered to be a term, while the same word in a documentary about the creation of the earth, is not.
2. Unit hood indicates that multi-word units should present a high degree of cohesiveness (1996) e.g. in *The earth machine* 'free-floating blocks of concrete' is considered to be one term and so is its corresponding term in Dutch 'vrij bewegende betonnen structuur'.
3. The last criterion is related to the translation itself. In order to compile a bilingual glossary for use by translators, all term pairs should be valid translation pairs (Macken et al., 2013). As the corpus for this study was provided by the Flemish public broadcaster, the translations are considered to be of an outstanding quality.

Matamala (2010) points out that science documentaries will always aim to popularise a specialised topic and to entertain the audience. She adds that science popularisation can be considered an instance of specialised discourse and that science documentaries, being specialised texts, contain terminological units.

Before starting the automatic terminology extraction, a quality test to measure the gold standard was carried out. First, three annotators labelled the terms of the second sub-corpus manually. Second, the inter-annotator agreement was calculated by means of precision, recall and F-score, the harmonic mean of precision and recall (Van Rijsbergen, 1979):

Precision = Number of correctly extracted terms / Total number of extracted terms

Recall = Number of correctly extracted terms / Total number of actual terms in the text

$$F - score = 2 * (precision * recall) / (precision + recall)$$

As the annotators do not always annotate the same set of items or terms, agreement measures such as kappa (Cohen, 1960) or alpha (Krippendorff, 1970) could not be used, and therefore F-score was chosen as agreement metric. F-scores are calculated by considering the annotations of one annotator as the

gold standard, and measuring precision and recall of the second annotator on that gold standard set of annotations. 'Precision gives an indication of whether the proposed terms are relevant, whereas recall measures the capability to retrieve *all* relevant terms in a given domain-specific document collection' (Macken et al., 2013). Table 1 lists the inter-annotator figures of the manual term extraction, starting from 724 aligned sentences for the three texts. The rates indicate that e.g. for *Madagascar* annotator one labelled only 24.86% of the terms labelled by annotator two whereas 71.32% of all terms labelled by annotator two were also labelled by annotator one. The output suggests a strong disagreement between annotator one and annotator two and three for *Madagascar* and in general, a fuzzy boundary between 'term' and 'non-term'. In *Madagascar*, for instance, one option is to limit terms to animals and plants, another to extend terms to all elements of nature. In *The earth machine*, terms might also include verbs related to geological phenomena. This is not a matter of 'right or wrong' but it provides evidence on how subjective the manual labelling can be. Whether one considers a word or a group of words as a terminological unit will inevitably remain a personal choice. However, for translational purposes a broad interpretation of what a term is, is more useful. Translators might want to check the bilingual glossary also in case of doubt about the spelling, to verify the accuracy of a translation or simply for inspiration. For this reason, the gold standard was created assembling the terms of all three annotators.

Table 1. Rate of agreement expressed in precision, recall and F-score between the 3 manual annotators.

Inter-annotator agreement in %		Madagascar	The Earth Machine	The Secret World of Pain
Ann.1	Precision	24.86	58.26	36.99
vs	Recall	71.32	68.89	48.80
Ann.2	**F-score**	**36.87**	**63.13**	**42.08**
Ann.1	Precision	23.68	43.18	43.37
vs	Recall	72.87	53.73	31.86
Ann.3	**F-score**	**35.74**	**47.88**	**36.73**
Ann.2	Precision	59.70	48.76	48.86
vs	Recall	64.05	51.30	47.35
Ann.3	**F-score**	**61.80**	**50.00**	**48.09**

Performance of the automatic terminology-extraction systems

The following step examined whether the terminology of the reference corpus is specific enough to be detected by automatic systems. The automatic terminology extraction of two commercial systems for

bilingual term extraction (SDL Multiterm Extract 2011 Trados® and Similis®) and one system being developed by the University of Ghent[2] (Macken et al., 2013) were tested. Sub-corpus two was used.

Macken (2010) explains that automatic terminology-extraction systems have basically

> 'two methodologically different approaches. The linguistic approach is based on the characteristics of term formation patterns, which are expressed as part-of-speech code sequences (e.g. N N, N prep N, Adj N). […] Linguistically-based terminology-extraction programs are always language dependent. [...] The statistical approach on the other hand is language independent and is based on quantifiable characteristics of term usage. One such characteristic is that terms tend to occur more frequently in specialized texts than in general domain texts' (p.105).

As pure statistically-based methods tend to over generate terms and pure linguistically-based methods produce some noise, most state-of-the-art systems use hybrid approaches. In 'Study and implementation of combined techniques for automatic extraction of terminology' Daille (1996) explores a method combining linguistic and statistical approaches using several statistical measures such as frequency, association scores, diversity and distance metrics.

The systems used in this study have different underlying technologies. SDL Multiterm Extract 2011 Trados® states on its website that the system is based on a statistical approach (http://www.translationzone.com/products/sdl-multiterm/extract/index-tab2.html#tabs) while Similis® is a hybrid system. It contains monolingual lexicons fed with glossaries extracted automatically from previous translations or imported from the translator's or the customer's own glossaries. The tool runs a linguistic analysis that sees sentences as a series of syntactical units called 'chunks', which are made up of a combination of words. The more interaction between the translator and the system, the better the result of the analysis (Planas, 2005). TExSIS is similar to SIMILIS®, generating first candidate terms from linguistically motivated aligned chunks, which are based on a shallow morphosyntactic automatic pre-processing of the texts. Both word alignment information and syntactic chunk information are taken into account for the creation of the bilingual candidate term list. The candidate terms are filtered by means of several filters for single-word and multi-word terms (Macken et al., 2013).

In the present research, the terminology lists resulting from the bilingual term extraction of the three systems were compared to the terminology lists of the GS, calculating the rate of agreement expressed in terms of precision in table 2 below. The higher the precision score, the more relevant the terms are. In other words, a higher rate of agreement means that more automatically extracted terms

[2] TExSIS: http://www.lt3.ugent.be/en/

correspond exactly to the terms labelled manually in the GS. As illustrated in table 2, TExSIS outperforms the two other systems for all three texts. These results are in line with those of Macken et al. (2013).

Table 2. Precision score for the bilingual terminology extraction of the GS and the three automatic term-extraction systems.

PRECISION in %	SIMILIS®	SDL MULTITERM EXTRACT®	TExSIS
The Earth Machine	28.78	25.37	58.67
Madagascar	33.33	32.08	55.97
The Secret World of Pain	24.32	21.43	53.66

In the tests Macken et al. (2013) conducted, the TExSIS precision scores were higher than ours (61.95 to 66.55) which can be explained by the audiovisual nature of our translations. The VRT guidelines explain that English off-screen dubbing translated into Dutch needs a free translation whereas for terminology extraction enough correspondence is needed between script and translation to be able to associate terms of the source text with their translation.

However, the bilingual glossary in table 3 demonstrates that both criteria, i.e. 'termhood' and 'unithood' are met. To measure the termhood criterion and to filter out general vocabulary words, TExSIS applies Log-Likelihood (LL) filters on all single-word terms (Rayson & Garside, 2000). In order to calculate LL, a frequency list is made for each corpus (the domain-specific and a background corpus). Then, log-likelihood is calculated for each word in the frequency lists. This is done by constructing a contingency table as is shown in table 3, where c represents the number of words in the first corpus, while d corresponds to the number of words in the second corpus. The values a and b are called the observed values (O).

Table 3. Contingency table to calculate Log-Likelihood

	First Corpus	Second Corpus	Total
Frequency of word	a	B	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	C	c+d

In the formula below, *N* corresponds to the total number of words in the corpus, *i* corresponds to the single words, whereas the "observed values" O_i correspond to the real frequency of a single word *i* in

the corpus. For each word *i*, the observed value O_i is used to calculate the expected value E_i according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

Applying this formula to our contingency table (with N1 = c and N2 = d) results in:

$$E_1 = c * (a + b)/(c + d)$$
$$E_2 = d * (a + b)/(c + d)$$

The resulting expected values can then be used for the calculation of the Log- Likelihood:

$$LL = 2 * ((a * log(\frac{a}{E_1})) + (b * log(\frac{b}{E_2})))$$

Words with a LL value above a predefined threshold are considered terms.

To measure the unithood the system calculates C-value for the multi-word terms. The C-value is an algorithm designed to recognise and extract multi-word terms (Frantzi & Ananiadou, 1996). It examines the string's frequency of occurrence in the corpus, its frequency of occurrence in longer candidate collocations and its length. The following formulas calculate the C-value for nested and non-nested candidate terms:

$C - Value(a) = \log_2|a|f(a)$ *if a is not nested*

$C - Value(a) = \log_2 \left|a\right|(f(a) - \frac{1}{P(Ta)} \sum_{b \epsilon Ta} f(b))$ *otherwise*

a = candidate term

b = a candidate term that contains *a*

$|a|$ = length of the candidate term (number of words)

f(a) = frequency of occurrence of *a* in the corpus

Ta = set of candidate terms that contain *a*

P (Ta) = number of candidate terms in *Ta*

f(b) = frequency of occurrence of *b* in the corpus

Longer collocations are considered more important than shorter appearing with the same frequency. The bigger the number of strings a substring appears in, the bigger the C-value of the string. C-value (a) = 0 means that (a) has the same frequency with a longer candidate collocation that contains (a) i.e. it is not a collocation.

The table below presents the first 52 terms of the bilingual glossary for *The earth machine* extracted by TExSIS, illustrating a term hood and unit hood.

Table 4. The first 52 candidate terms generated by TExSIS.

English	Dutch	Termhood (LL)	C-value
crust	aardkorst	29.15	0.03144
crust	korst	29.15	0.03144
earth	aarde	35.41	0.05393
earth	aardkorst	35.41	0.05393
earth	aardoppervlak	35.41	0.05393
super-continent		3.77	0.00144
lava	lava	20.23	0.02134
lava	vulkaan	20.23	0.02134
mantle	mantel	12.38	0.01442
mantle	buitenmantel	12.38	0.01442
planet	aarde	16.92	0.02660
planet	planeet	16.92	0.02660
planet	oceanen	16.92	0.02660
tectonic		11.84	0.00901
surface	aardoppervlak	15.84	0.03244
surface	grond	15.84	0.03244
surface	oppervlakte	15.84	0.03244
volcanoes	vulkanen	9.31	0.01009
fault	breuklijn	11.46	0.01550
fault	verschuiving	11.46	0.01550
fault	breuk	11.46	0.01550
miles	kilometer	12.58	0.03317
miles	meter	12.58	0.03317
plates	platen	9.26	0.01346
crater	krater	7.39	0.00865
crater	kraterwand	7.39	0.00865

smoke-shrouded rim		0.00	100.072
heat	hitte	9.99	0.02019
heat	warmte	9.99	0.02019
fourty-five miles		0.00	100.072
five-and-half miles		0.00	100.072
storm-tossed seas	woelig water	0.00	100.072
never-ending movement		0.00	100.072
tectonic plates	tektonische platen	0.00	500.361
tectonic plates	platen	0.00	500.361
All-year-round swimming		0.00	100.072
molten		5.72	0.00469
stadium	stadion	6.92	0.01009
earthquake	aardbeving	5.74	0.00865
earthquake	aardbevingen	5.74	0.00865
aquifer	aquifer	9.16	0.00433
earthquakes	aardbevingen	5.96	0.00577
hangar		4.51	0.00577
cracks	scheuren	5.21	0.00721
cracks	scheurtjes	5.21	0.00721
rock	gesteente	7.27	0.02019
rock	rotsen	7.27	0.02019
turtles	schildpadden	4.69	0.00577
turtles	tank	4.69	0.00577
super-heated molten rock		0.00	158.544
magnetic field	magnetisch veld	0.00	900.649
magnetic field	veld	0.00	900.649

Experiments

It is now possible to address research question three: Does an automatically extracted bilingual glossary help to translate documentary texts as far as process time and number of terminological errors are concerned? Twelve master students with average to good marks, enrolled in an English-Dutch translation course, participated in the proof-of-concept translation experiments. The set-up of these experiments was designed as follows:

- The source text was based on sub-corpus two, the same three episodes that served as a basis for the term extraction. A selection of sentences was made from the off-screen dubbing of the English script for a total amount of 775 words. Each sentence contained one or more different terms. To ensure a correct comprehension of the text, additional

information was added where needed between brackets. With 'term' we mean 'labelled as such in the GS'. Table 5 below shows the number of term types of the source texts listed in the two glossaries and indicates that only one term was extracted by TX but not labelled in the GS.

Table 5. Number of term types per text in the glossaries.

Number of term types	only in GS	only in TX	GS and TX
Earth Machine	16	1	23
Madagascar	31	0	17
Pain	34	0	16

- The participants did not make use of the video as the source text consisted of separate fragments with no coherence between them. Moreover, the terminology itself in the text was clear and a particular timing or style matching with the images and the voice talent was not required for the purpose of this study. Furthermore, in some cases, for rush jobs and/or for copyright reasons, audiovisual translators have no access to the video.
- All the participants translated the same source text from English into Dutch.
- In a first session, all the participants translated without bilingual glossary but they were allowed to consult all the digital sources they wanted to, including dictionaries available on the university website.
- In a second session two months later (considered to be the time needed to 'forget' the first translation), the same participants were divided into two groups. In order to guarantee the same level of competence in both groups, the translation made for session one has been assessed so that two comparable groups could be made. The first group was asked to translate the same text with the GS, the second group translated with the bilingual glossary extracted automatically by TExSIS.
- While the students were translating, Inputlog[3] – a keystroke logging tool that logs all types of input modes, being developed at the University of Antwerp – registered the whole translation process. By the means of this tool, an analysis of the total translation time, pauses before terms and search for information was possible (Leijten & Van Waes, 2013).

A trial run with one student tested the technical and organisational efficacy of this set-up. The time needed for the translation of the source text was calculated to be two and a half hours. The source text

[3] http://www.inputlog.net/

was copied onto each desktop in three MS Word-documents (*The earth machine, Madagascar* and *The secret world of pain*) as were the glossaries in the case of the second session; the bilingual English-Dutch-English dictionary, the monolingual Dutch dictionary and the Google homepage were accessible.

The students were asked to translate each sentence into a definitive version and to do revisions only while translating that sentence. Post-editing would have complicated a correct tracking of the pauses before terms. Each target text was logged in a separate file so data loss because of technical failures would be limited to maximum one text. The participants were told to bear in mind the time limit (2h30) but not to rush in order to finish, which would have reduced the pause time for terms.

Due to a human and a technical failure, two target texts were not logged. However, the translation itself was saved in both cases, so it could be used for the analysis of the terminological errors and for the assessment of the translation competence.

Data analysis

Process time and pause time

Using the summary and the general analysis that Inputlog provides, the complete translation process was analysed. First, the total process time of all translations, displayed in milliseconds in each Inputlog summary analysis, was calculated. The results in Table 6 confirm the hypothesis that the integration of bilingual, domain-specific glossaries increases the efficiency of the translation process. When working without glossary, only half of the students were able to finish the three texts. Six students translated only 31 to 64% of the last text. For this reason, the last text was not included in the statistical analyses.

Table 6. Average process time rounded off to the nearest second for both groups in both conditions.

Average process time in sec	TX group	GS group
With glossary	2494	2076
Without glossary	3018	2997

The difference between the total process time in the two conditions was examined. The data were analysed in SPSS version 20. A two samples paired t test was used, as all assumptions for using this test were satisfied. On average, the participants spent significantly more time to translate the two texts when working without glossary than when working with glossary. The average process time difference for the TExSIS group for both texts was 524.29 s (C.I. of the difference 259.30 to 789.29 s, $p = .001$) and for the gold standard group 920.92 s (C.I. of the difference for 496.22 to 1345.63 s, $p = .001$).

As this research is interested especially in terminology, the focus was then narrowed to the pause time before terms labelled manually as such in the gold standard. Inputlog's general analysis yields the very detailed analysis of the writing process needed for this type of study. Leijten and Van Waes (2013) explain that the output features represent every log event, the cursor position, the document length, the start and end time of every event in milliseconds and are used to calculate the action and pause times. In order to determine the pause times before terms, we scrolled the log files manually selecting the rows belonging clearly to a pause time before terms i.e. the event from the moment the participant enters a dictionary, the internet or the bilingual glossary until he/she enters the 'Wordlog' document in which the translation is written down. If the participant surfs from one source to another without going to the 'Wordlog' document in between, e.g. first to the dictionary then to the internet, this was considered one pause time. Table 7 below shows a significant difference between the average pause time for terms in the group translating without and with the glossary for both groups.

Table 7. Average pause time for terms rounded off to the nearest second for both groups in both conditions.

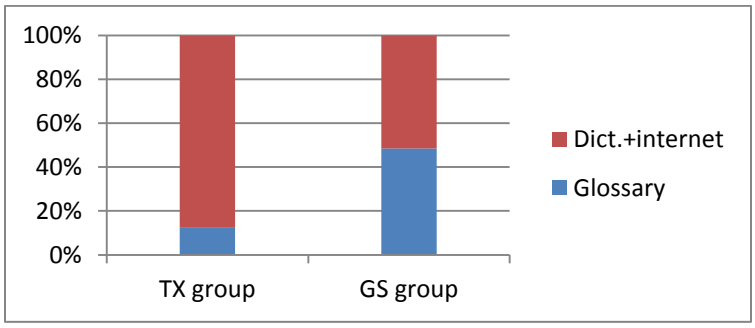
Average pause time for terms in sec	TX group	GS group
With glossary	443	276
Without glossary	710	613

The difference between the pause time in the groups working without and with glossary was addressed. The data were analysed in SPSS version 20. A two samples paired t test was used, as all assumptions for using this test were satisfied. On average, participants spent significantly more time to translate the terms in the first condition (without glossary) than in the second condition. The average pause time for terms for the TExSIS group was 267.17 s (C.I. of the difference for : 154.54 to 379.81, p = .000) and for the gold standard group 337.79 s (C.I. of the difference for 156.45 to 519.13, p = .002). Comparing the average pause time between the two conditions in table 7, we can observe a pause time gain for terminology of 38% for the TExSIS group and 55% for the gold standard group, whereas a comparison between the average process time in the two conditions (table 6) indicates a process time gain of 17% for the TExSIS group and 31% for the gold standard group.

In an attempt to understand the importance of a bilingual glossary, an analysis of the use of sources was made. The total pause time before terms was divided into two categories: pause time for bilingual glossaries (GS or TExSIS) and pause time for online dictionaries plus the internet (i.e. all other internet sources the participants consulted to gather information and find the correct translation). The graph below (figure 1) presents how the two categories were spread over the total pause time before terms. In the case of the gold standard group, the time for consulting the glossary and the time for the

dictionaries plus the internet is equally distributed, whereas the TExSIS group spent clearly much more time consulting dictionaries or the internet. In other words, the more exhaustive the glossary is, the less time the translators spent scrolling through the internet and in dictionaries.

Figure 1. Use of the glossaries vs use of dictionaries and internet per group

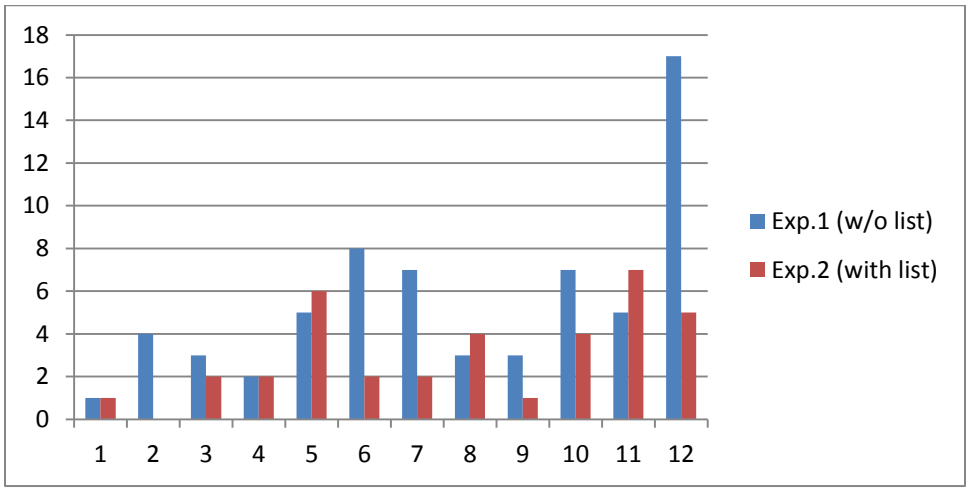


Terminological errors.

As a measure of the impact of the glossaries on the translation quality, all terminological errors were assessed. A terminological error means that a term from the gold standard was translated with a term not corresponding to a correct translation in this context.

The graph below (figure 2) provides data on two source texts for a total of 456 words (the third text, which half of the participants did not complete in the first session, was not considered). Seven out of twelve students made fewer errors working with the list, two did equally well and three did worse.

Figure 2. The n° of terminological errors per student for the translation of 456 words without and with glossary.



The corresponding paired t test showed no significant difference in the number of errors between the two conditions (p= .06) considering the data of both the GS group and the TX group. However, analysing the number of errors of the GS group only, translating without and with the glossary, a significant

difference can be observed. The data were analysed in SPSS version 20. A two samples paired t test was used, as all assumptions for using this test were satisfied. On average, participants made significantly fewer terminological errors working with the GS than working without. The average number of errors is 2.16 (C.I. of the difference for : 0.45 to 3.87, p = .01).

Discussion

The first research question of this article was confirmed: we have found that an exhaustive, manually labelled bilingual glossary supports inexperienced translators in terms of process time and terminological errors. Indeed, both the process time translating with glossary and the number of terminological errors improve significantly when the students work with the gold standard. Next, the study investigated whether the terminology in documentaries is specific enough to be detected by automatic terminology-extraction systems and found that the bilingual glossary automatically extracted by TExSIS meets the criteria of term hood as well as unit hood, although its precision rate shows room for improvement. As for the third research question and the third step in this study, i.e. the support an automatically extracted bilingual glossary provides for the translation of documentaries, this was shown to reduce the process time of the participants using the TExSIS glossary significantly but not the number of terminological errors. Unfortunately, we have insufficient data to elaborate a statistical analysis of the difference in errors between the GS group and the TX group. Moreover, as none of the participants worked with both glossaries it is not possible to explore the correlation between the exhaustiveness of two different glossaries and the number of term errors in the target text.

Interestingly, in 31% of the cases the glossary did not help to improve the translation, the errors made in session one were not corrected in session two in spite of the glossary. The participants do consult the glossary but they do not rely on it. They double check the term in a dictionary and/or through the internet and then make their (wrong) translation choice. This lack of trust might be a beginners attitude, used as they are to rely heavily on dictionaries and on the internet without the experience to recognize the correct translation choice through a high amount of information. It can be assumed that the longer translators work with bilingual glossaries, the more they learn to rely on them. Yet, with only three texts at our disposal, the sample of this experiment was too small and the translation time too short to deduct a change in attitude towards the glossary.

In conclusion, we can argue that the integration of a domain-specific terminology glossary increases the efficiency of the translation process for unexperienced translators. Expanding the model of this pilot study, which served to verify the validity of the set-up and methodology, to professional

translators has been the next step in this research and it was completed by the time of submission of this article. The results will be presented in a follow-up publication. We trust that we will be able to provide a useful impetus for the use of terminology-extraction tools for the translation of documentaries and contribute in this way to a more efficient translation process when dealing with terminological challenges in this type of film. Such an achievement is particularly relevant at this moment: the audiovisual translation market operates at very competitive prices and time is money more than ever. Applying translation support tools might allow translators to reinforce their position on the translation market.

Aknowledgements

The subject of this research was inspired by Anna Matamala's project for the Spanish Ministerio de Economia y Competitividad, reference code FFI2012-31024, 'Accesibilidad lingüìstica y sensorial: tecnologìas para las voces superpuestas y la audiodescripciòn'.

References

Bibliographical references

Christensen, T., & Schjoldager, A. (2010). Translation-memory (TM) research: What do we know and how do we know it? *Journal of language and communication studies, 44,* 89-101. Retrieved from http://download2.hermes.asb.dk/archive/2010/Hermes44.html

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* (1), 37-46. doi: 10.1177/001316446002000104

Coombs, J. (2014). How a large biotechnology company teamed with a translation service provider to define best practices. *Journal of commercial biotechnology, 20* (1), 49-53. doi: 10.5912/jcb638

Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. L. Klavans and P. Resnik (Eds.), *The balancing art: Combining symbolic and statistical approaches to language* (pp. 49-66). Cambridge MA: MIT Press.

Franco, E. (2000). Documentary film translation: A specific practice? In A. Chesterman, N. Gallardo San Salvador and Y. Gambier (Eds.), *Selected Contributions from the EST Congress Granada 1998: Translation in Context* (pp. 233-242). Amsterdam: John Benjamins.

Franco, E., Matamala, A. & Orero, P. (2010). *Voice-over translation: An overview***.** Bern**:** Peter Lang.

Espasa, E. (2004). Myths about documentary translation. In P. Orero (Ed.), *Topics in Audiovisual Translation* (pp. 183-197). Amsterdam/Philadelphia: John Benjamins.

Frantzi, K. & Ananiadou, S. (1996). Extracting nested collocations. *Proceedings of the 16th conference on computational linguistics, 1,* 41-46*.* doi:10.3115/992628.992639

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition. A review. *Terminology, 3,* 259-289*.* doi: 10.1075/term.3.2.03kag

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement, 30* (1), 61-70. doi: 10.1177/001316447003000105

Lagoudaki, E. (2010). *Translation memories survey 2006. Translation memory systems: Enlightening users' perspective.* Paper presented at ASLIB International Conference on Translating and the Computer. 2006, London. Retrieved from http://mt-archive.info/Aslib-2006-Lagoudaki.pdf

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30* (3), 358-392. doi: 10.1177/0741088313491692

Le Blanc, M. (2013). Translators on translation memory (TM). The results of an ethnographic study in three translation services and agencies. *Translation & Interpreting, 5* (2), 1-13*.* doi: ti.105202.2013.a01

Macken, L. (2010). *Sub-sentential alignment of translational correspondences* (Doctoral dissertation). Retrieved from https://biblio.ugent.be/input/download?func=downloadFile&recordOId=966953&fileOId=966963

Macken, L., Lefever, E., & Hoste, V. (2013). TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology*, *19* (1), 1-30. doi 10.1075/term.19.1.01mac

Matamala, A. (2009). Translating documentaries: From Neanderthals to the Supernanny. *Perspectives*, *17*, 93-107. doi: 10.1080/09076760902940112

Matamala, A. (2010). Terminological challenges in the translation of science documentaries: A case-study. *Across Languages and Cultures, 11,* 255-272. doi: 10.1556/Acr.11.2010.2.7

Osstyn, K. (2010). *VRT Leidraad voor commentaarvertalen 2010.* Unpublished manuscript.

Planas, E. (2005). *Similis. Translation memory software.* Paper presented at ASLIB International Conference on Translating and the Computer. 2005, London*.* Retrieved from http://mt-archive.info/Aslib-2005-Planas.pdf

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the workshop on comparing corpora, 9*, 1–6. doi: 10.3115/1117729.1117730

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths.

Wright, S. E. (1997). Term selection: The initial phase of terminology management. In S. E. Wright and G. Budin (Eds.), *Handbook of terminology management* (pp. 13-23). Amsterdam: John Benjamins.

Filmography

Sub-corpus one

Monty, D. (Writer), Finch, T. (Producer). (2008). Around the world in 80 gardens – Australia/New Zealand – Episode 2 [Television series episode]. In K. Richardson (Series Producer), *Around the world in 80 gardens.* UK: BBC2.

Chapman, A., Jackson, C. & Smits van Oyen, M. (Producers). (2006). Big cat week – Series 3, Programme 1 [Television series episode]. In S. Ford [Executive Producer], *Big cat week.* UK: BBC.

Gates, S. (Writer) & Smith, S., Williamson, T. (Directors). (2010). E-numbers, an edible adventure - episode 1 [Television series episode]. *E-numbers, an edible adventure.* UK: Plum Pictures Production Ltd/BBC Worldwide.

Gyves, M., Learoyd, S. (Producers), & Gyves, M. (Director). (2010). How earth made us – Deep earth - Episode 1. [Television series episode]. In J. Renouf (Series Producer), *How earth made us.* UK: BBC/National Geographic.

Nicopoulos, S. (Writer), & Rose, Y. (Director). (2010). *Is the magnetic pole about to flip?* [Motion picture]. Canada: TGA Production/Ideacom International Inc./CNRS Images .

Summerill, M. (Producer). (2011). Madagascar – Island of marvels [Television series episode]. In M. Gunton (Executive Producer), *Madagascar.* UK: BBC NHU/Animal Planet.

Dalton, Ph. (Producer), & Downer, J. (Director). (2011). Polar bear – Spy on the ice - 1 [Television series episode]. In C. Harrison (Executive Producer), *Polar bear – Spy on the ice.* UK: John Downer Production Ltd for BBC Worldwide/Animal Planet.

Haken, L., Warner, L. & Roberts, R. (Producers), & Shoolingin-Jordan, N., Slee, M. (Directors). (2011). Earth machine - Land [Television series episode]. In L. Van Beeck (Series Producer), *Earth machine*. UK: BBC/Discovery Channel.

Gray, I. (Producer), & Brown, J. (Director). (2010). The natural world – Empire of the desert ants - Season 31 – episode 2 [Television series episode]. In T. Martin (Executive Producer), *The natural world.* UK: BBC/Animal Planet.

Learoyd, S. (Producer & Director). (2011). Horizon - The secret world of pain – Season 47 – episode 11 [Television series episode]. In A. Laverty (Series Producer), *Horizon.* UK: BBC2.

Sub-corpus two

Haken, L., Warner, L. & Roberts, R. (Producers), & Shoolingin-Jordan, N., Slee, M. (Directors). (2011). Earth machine - Land [Television series episode]. In L. Van Beeck (Series Producer), *Earth machine*. UK: BBC/Discovery Channel.

Learoyd, S. (Producer & Director). (2011). Horizon - The secret world of pain – Season 47 – episode 11 [Television series episode]. In A. Laverty (Series Producer), *Horizon.* UK: BBC2.

Summerill, M. (Producer). (2011). Madagascar – Island of marvels [Television series episode]. In M. Gunton (Executive Producer), *Madagascar.* UK: BBC NHU/Animal Planet.