To my parents

Promotoren	Prof. dr. Véronique Hoste Vakgroep Vertalen, tolken en communicatie - Universiteit Gent Prof. dr. Aline Remael Vakgroep Vertalen en tolken - Universiteit Antwerpen
Decaan	Prof. dr. Marc Boone
Rector	Prof. dr. Anne De Paepe

Nederlandse vertaling:
Het vertalen van documentaires: helpt een bilinguaal glossarium van domein-specifieke terminologie om het vertaalproces te verlichten?

Kaftinformatie: Foto cover: ©Trui Hanoulle

FACULTEIT LETTEREN
EN WIJSBEGEERTE

Translating documentaries

Does the integration of a bilingual glossary of domain-specific terminology into the translation process reduce the translators' workload?

Sabien Hanoulle

Proefschrift voorgelegd tot het behalen van de graad van Doctor in de vertaalwetenschap

2017

UNIVERSITEIT
GENT

Tradurre è trovare la nota giusta
Umberto Eco

Acknowledgements

First and foremost, I want to thank my supervisors, Aline Remael and Véronique Hoste, for their constant guidance and support throughout this enlightening project. Despite demanding schedules of their own, they were always there with expert advice, clear and to the point, with refreshing ideas when I got stuck and more importantly, with words of encouragement when I was about to give up. They were an invaluable point of reference, and it is safe to say that I could not have finished this thesis without them.

I gratefully acknowledge Anna Matamala for suggesting the subject of this dissertation when I discovered that a book had just been published about the very project I was working on. Even though our research did not turn out to be a joint project after all, this experience has been such a learning curve for me. Anna introduced me to the intriguing world of (translating) documentaries, which I consider a lifelong enrichment.

I also wish to thank the Flemish broadcasting corporation, VRT, which generously placed all the documentaries they broadcast between 2005 and 2013 at my disposal. I selected my corpus from those documentaries, one of the key components of this research.

Another component of utmost importance are the experiments. Twelve students and twelve professional translators gave hours of their precious time and energy to help me carry out the most crucial part of the project. Without their availability, effort and knowledge, this thesis would not have been possible. Thank you all.

I also want to acknowledge Eric Van Horenbeeck from the University of Antwerp who introduced me to Inputlog, the keystroke logging tool used for the experiments, LT3-members Lieve Macken and Els Lefever for doing the term extraction with TExSIS, Peter Velaerts for helping me with the technical issues of TExSIS, Marjan Van de Kauter for initiating me into manual term extraction and Kristien Temperville for proofreading my final text, without even knowing her in person.

Very special thanks also go to my uncle Ignace Hanoulle who helped with the statistical analyses and managed to impart the basics of statistics with never ending patience. His proposals and our discussions opened new research paths to be explored.

My colleagues Christophe, Gert and Isabelle deserve a special mention: Christophe for taking over duties in spite of his double appointment in London and in Antwerp, Gert for changing his mind and Isabelle for accepting the 'back-up position'. I have many great colleagues though, who shared difficult moments and enquired about my progress. Anyone who has written a PhD will know how heartening solidarity among colleagues can be.

I also have many great friends who never stopped believing that I would bring this long-term task to an end. They were always understanding when I disappeared behind my books instead of joining them for a coffee in the 'Clouds', a dinner or a chat. Sorry for neglecting you, I promise I will make it up to you.

Zia Paola, Franca e Lorenzo, grazie per esservi presi cura di Matilde, tanti (troppi) estati, vacanze di Natale e Pasqua, senza mai lamentarvi del peso, anzi, sempre a braccia aperte. Cosi, io potevo continuare a lavorare, sapendo che lei era in buone mani.

To my parents, sister, brother and sister-in-law, I am deeply grateful to you for always being there – even from a distance – supporting me, as you did, with your family warmth, for believing in me, no matter what. So proud to be yours and so proud you are mine!

Last but not least, I want to thank my daughter Matilde, for reminding me how important it is to sit on the couch and chill. My husband Andrea, I owe most of all. Thanks for not complaining (too much) about all those nights I spent behind my computer. It is thanks to your patience and understanding that we are still a family and we can now pick up where we left off.

Abstract

The thesis under study is a contribution to the research in translation of documentaries and more particularly, the translation of domain-specific terminology, one of the challenges in this field. Carrying out translation experiments, this dissertation investigates whether or not bilingual glossaries, manually or automatically extracted, reduce the workload of documentary translators.

The research consists of three major parts, all thoroughly analysed against the state-of-the-art studies. The first part concerns the analysis and the selection of the corpus. The Flemish public broadcaster VRT made a corpus of English documentaries and their Dutch translation available. In a preparatory stage, an in-depth analysis of the text type showed the corpus contained domain-specific terminology, especially in documentaries meant for informative purposes. As a consequence, an experimental corpus of three informative documentaries was selected for the translation experiments.

The second part focuses on manual and automatic terminology extraction, the underlying software of automatic term extractors and the testing of three existing systems. In order to understand the test results, two key features for terminology (termhood and unithood) were discussed and an overview of the different strategies term extractors use was provided. Annotators manually labelled all the terminology of the experimental corpus, drawing up in this way a gold standard as an objective means for testing the automatic systems. The accuracy of these systems was expressed in terms of precision and recall. The best performing system was selected to extract the glossaries for the experiments.

The third part deals with the translation experiments. In a pilot project, Master's students in translation translated three texts first without, then with the manual or the automatic glossaries at their disposal, while a keystroke logging software registered the process. For the main experiment with professional translators, the experimental set-up was slightly modified, introducing some remedial measures learned from the pilot project. Statistical analyses of the total process time and the pause time before each term were elaborated.
The results revealed that in most working conditions the candidates worked significantly faster when translating with a glossary and that they made less terminological errors. Furthermore, the dissertation proposes an ecologically valid experimental design, tested and remediated in the ongoing research. Yet, there was room for improvement for the automatically extracted glossaries due to the small corpus and the free translation style, typical for translating documentaries.

Samenvatting

Dit proefschrift levert een bijdrage aan het onderzoek naar de vertaling van documentaires, en met name naar de vertaling van domeinspecifieke terminologie, een van de uitdagingen van dit genre. Experimenten onderzochten of de werklast van vertalers daalt als zij documentaires vertalen met manueel of automatisch geëxtraheerde bilinguale glossaria.
De studie bestaat uit drie delen die grondig getoetst zijn aan bestaande onderzoeken. Het eerste deel behandelt de analyse en de selectie van het corpus, dat ter beschikking was gesteld door de Vlaamse publieke omroep VRT. Een voorbereidende fase toonde aan, met een gedetailleerde analyse van de tekstsoort, dat het corpus domeinspecifieke terminologie bevat, vooral bij documentaire films voor informatieve doeleinden. Het experimentele corpus van drie documentaires werd dan ook geselecteerd uit de informatieve films.

Het tweede deel omvat een studie van manuele en automatische termextractie, de onderliggende software van automatische termextractiesystemen en een test van drie bestaande termextractors. Om de resultaten van de test te kunnen interpreteren zijn twee cruciale factoren van terminologie (termhood en unithood) besproken en is een overzicht gegeven van de verschillende strategieën waarop termextractiesystemen gebaseerd zijn. Annotatoren hebben een gouden standaard van manueel gelabelde termen opgesteld waarmee drie termextractors werden getest. Aan de hand van 'precisie' en 'recall', die de accuraatheid van de systemen uitdrukken in cijfers, kon het beste systeem worden uitgekozen om de bilinguale glossaria voor de experimenten te extraheren.

Het derde deel beschrijft de experimenten. In een pilootstudie vertaalden masterstudenten drie teksten, eerst zonder, dan met manueel of automatisch geëxtraheerde glossaria die ze ter beschikking hadden, terwijl een keylogger software het proces registreerde. Het hoofdexperiment met professionele vertalers werd op dezelfde manier uitgevoerd, behalve enkele remediërende wijzigingen die noodzakelijk bleken uit de pilootstudie. Met statistische analyses werden de totale procestijden en de pauzetijden voor elke term verwerkt.

Uit de resultaten bleek dat in de meeste werkcondities de deelnemers significant sneller werkten met glossaria dan zonder en dat ze minder terminologische fouten maakten. Het onderzoek heeft tevens een ecologisch valide experimenteel design uitgewerkt. De typisch vrije vertaling van documentaire teksten en het kleine corpus maakten echter dat de automatische glossaria voor verbetering vatbaar zijn.

List of Tables

List of Figures

Table of Contents

Introduction

Translation is rapidly evolving into a semi-automated working process thanks to thorough research in fields like computer science, linguistics and, of course, translation studies. Electronic dictionaries, the Internet with its numerous possibilities for research and communication, computer-assisted translation tools and machine translation have profoundly changed the translators' working conditions during the last decades. These computerised systems are meant to enhance the efficiency, speed and quality of translation tasks (Amparo, 2008). In the meanwhile, the market operates at very competitive prices and time is money more than ever. Applying translation support tools allows translators to reinforce their position on the market.

The present thesis is a contribution to research concerning translation support tools covering both translation studies and language technology. It focusses on audiovisual translation and more particularly on the translation of documentaries. Espasa (2004) points out that translating documentaries presents specific challenges, one of which is domain-specific terminology. Whereas for technical translations, it is common practice to organise terminology into databases by means of automatic terminology extraction systems, for translating documentaries, terminology is usually treated manually.

The audiovisual translators working for VRT, the Flemish public broadcaster who provided the corpus for the thesis under study, do not use CAT-tools nor automatic term extraction systems (G. Saerens, head of the translation departement at VRT, personal communication, June 7, 2017), nor do the six audiovisual translators who participated in the experiments conducted for this research. The only audiovisual translation mode at VRT where translation technology is actively applied is live subtitling. Carried out in real time, for live broadcasts, it is commonly used for intralingual transfer. The subtitler repeats or rephrases what is said on screen and the speech recognition system 'translates' the short utterences into written lines (Gambier, 2012). In order to avoid errors and to speed up the process, the system is fed with manual terminology lists, containing domain-specific terminology, named entities and all kinds of recurring words the system might fail to recognize.

However, several research projects have been carried out in order to introduce translation technology into different kinds of audiovisual translation (see Section 1.2.4). One of these is the STON project[1], conducted by VRT, two academic partners and three companies[2], aiming at the automation of intralingual subtitling of documentaries and newscontent for the deaf and the hard of hearing. No final results have been published yet (M. Lycke, researcher at VRT Innovatie, personal communication, June 6, 2017).

Similarly to other translation tasks, also the interlingual translation of documentaries might benefit from a mechanism "to avoid duplication of efforts" (Matamala, 2010, p. 269). Such a mechanism could for example be a terminological support by means of a terminology extraction system. As audiovisual translators of documentaries are often confronted with domain-specific terms, I hypothesize that it must be beneficial to translation efficacy to support the translators in their work with a domain-specific, bilingual glossary. The title of this dissertation formulates this hypothesis in a main research question. To test this hypothesis translation experiments were conducted in which candidates translated a documentary text with and without manual or automatic bilingual glossaries to assess whether or not there is a significant difference between working with and working without a glossary.

The first step to carry out these experiments, is the selection of the corpus. The Flemish public broadcaster VRT, made available eight years (2005-2013) of English documentaries and the Dutch script translation into subtitles and off-screen dubbing, the two translation modes used in Flanders. Off-screen dubbing can be defined as the translation of the commentary narration in documentaries, when on-screen speakers and the original audio tracks are absent (see also Chapter 1, Section 2). As (the automation of) subtitling has been the subject of many studies, the present research focusses on off-screen dubbing.

In order to determine the density of domain-specific terminology in the data, a preparatory corpus of ten documentaries about earth and space, wildlife and human body was selected and its terms were manually labelled and analysed. The specificity and number of terms revealed to be tailored to the subject and the target audience. Out of this preparatory corpus, an experimental corpus of three representative documentaries was selected.

---

[1] The STON project was funded by VRT and IWT-Innovatief Aanbesteden.
http://innovatie.vrt.be/home/2017/1/19/ston-aan-het-woord (accessed 17 June 2017).

[2] The universitiyies of Ghent and Leuven and Devoteam, Limecraft and PerVoice.

In a second step, the domain-specific, bilingual glossaries for the experiments were created. Labelling manually all terms, a gold standard was drawn up as an objective means for testing three automatic term extraction systems, viz. SDL Multiterm Extract 2011 Trados®, Similis® and TExSIS[3]. Precision and recall measures illustrated their accuracy and the best performing system was selected to extract the automatic glossaries.

The third step consisted of the translation experiments. A pilot study with Master's students in translation served as a proof-of-concept. The students were divided into two groups, one group working with the gold standard, the other one with the automatic glossaries. They all translated three texts, in the first session without glossaries, in the second session with glossaries. Inputlog, a keystroke logging software developed by the University of Antwerp, registered the writing process, providing all necessary data for statistical analyses.

Next, the same experiment was conducted with professional translators, applying a slightly different experimental set-up, which resulted from the pilot study. Thanks to this remedial measures, an ecologically valid[4] experimental design was obtained.

Calculating the total process time and the pause time before each term, statistical analyses were carried out in order to investigate the difference between translating with and translating without glossaries. In addition, the use of the information sources (dictionaries, internet, glossaries) and the terminological errors with and without glossaries were examined. A short retrospective survey completed the findings.

The thesis is outlined in four chapters, a discussion and a conclusion.

Chapter 1 offers a literature review regarding the three fields this dissertation includes: documentary, audiovisual translation and terminology extraction studies.

Section 1 explores the field of the corpus viz. documentaries from the point of view of film studies. An analysis of what is to be considered as a documentary is carried out, followed by a classification of the genre, its new trends, the verbal mode in documentaries and screenplay narration techniques.

Section 2 provides insights into the research about documentaries in audiovisual translation studies. The specific challenges of translating documentaries are discussed in detail before narrowing the focus on verbal narration and its lexical features. A combination of film studies with research in linguistics and in translation studies illustrates that commentary narration is the dominant verbal narration technique for informing. Furthermore, an analysis of Matamala (2010) and Cabré (2003) provides support for the view that science and nature documentaries contain domain-specific

[3] Developed by the University of Ghent.

[4] '[L]a légitimité de l'extrapolation au monde réel à partir des résultats obtenus dans l'environnement de recherche, notamment expérimental [...]' (Gile, 2011, p. 48).

terminology. Section 2 concludes with an overview of existing research projects in the automation of audiovisual content.
Section 3 addresses the main aspects of terminology and terminology extraction studies. Definitions of terminology together with their basic elements 'concept' and 'term' are discussed and the importance of terminology management is highlighted through an overview of the research into this field. Next, different strategies for terminology extraction are explained. As translation is a crucial issue of this dissertation, three bilingual term extraction systems are proposed in order to test their performance with the corpus under study.

Chapter 2 reports on the corpus of documentaries made available by VRT. The language pair, the translation mode as well as the subjects are discussed. A preparatory corpus of ten documentaries is selected in order to investigate whether or not it contains terminology. Out of this preparatory corpus, an experimental corpus is constituted in order to carry out the experiments. Moreover, a detailed analysis of the translation 'style' is carried out, revealing interesting aspects that affect the automatic term extraction.

Chapter 3 describes the methodology used to answer the main research question. The preparatory stage includes sentence alignment and manually labelling of the terms in the preparatory corpus, followed by the selection of the experimental corpus. Next, I report on the creation of the gold standard and the testing of three automatic term extraction systems. The chapter also provides more details about the methods for the experiments and the keystroke logging software applied to register the translation process.

Chapter 4 represents the core of this thesis: the translation experiments. It focusses first on statistical filters used by automatic term extractors, offering important background information to understand the accuracy of the automatically extracted glossaries. The experimental set-up of both the students' and the professionals' experiment is addressed, as well as the data analyses and the results. Considerable attention is devoted to the remediation of the experimental design.

Chapter 5 and 6 conclude this thesis with a discussion and a conclusion, proposing also suggestions for future research.

Chapter 1
LITERATURE REVIEW

Since the research question investigated in this dissertation contains three main elements, viz. documentaries, translation and bilingual glossaries of domain-specific terminology, the literature review explores the three research fields covering these elements.

1.1 Insights from film studies

The first element, documentaries, constitutes the 'text' type the thesis under study investigates: an audiovisual product with certain characteristics. This section discusses what is to be considered as a documentary and then narrows down the focus to the features involving terminology viz. classification and verbal mode in documentaries.

1.1.1 Origins of the term 'documentary'

While giving an overview of the history of documentary is beyond the scope of the present dissertation, a brief historical introduction is needed to better understand the concept of documentary film and its ambiguity. A documentary is a representation of actuality but also includes a creative aspect (see Section 1.1.2). Furthermore, every representation being a transformation (Spence & Navarro, 2011) is an important issue in the present thesis because the artistic ambitions of a documentary and the degree of transformation it entails, can affect the amount of terminology (still) present. Section 1.2.1 elaborates on the special challenges deriving from this hybrid characteristics of the genre and considers a few common translation strategies.

In addition, the genesis of a documentary also shapes its final form and must be considered when evaluating the features of the different subgenres and their degree of specialisation. This also has implication for the selection of productions that will (hypothetically) contain the most domain-specific terminology.

Over the last twenty years, the documentary has returned to the big screen. A new wave of films with many different subjects from countries all over the world has entered the cinema in a small but growing number. Documentaries should now be considered a complex genre, multifaceted and influenced by a range of different contexts. The digital revolution has changed the technical and financial accessibility to information, to exhibition platforms and to distribution channels for both audience and producers (Chanan, 2007; Chapman, 2009).

The discussion about what is to be considered a documentary and what not, however, is an ongoing issue. Many scholars like Nichols (1991), Plantinga (1997), Renov (1993), Ward (2005) and Winston (1995), to name but a few, focussed on the difference between fiction and non-fiction to define the term. Others, in more recent years like Chanan (2007) and Chapman (2009), broadened the concept, offering a different approach. I will elaborate on this discussion after a brief introduction to the genesis of documentary.

The need for documenting a phenomenon or an action drove scientists and photographers in the late 19th century to experiment with cameras. The French astronomer Janssen recorded Venus passing across the sun in 1874 and at about the same time, the English photographer Muybridge, sponsored by an American horse-breeder interested in improving speed, registered a galloping horse (Barnouw, 1993).

Although these experiments did not yet result in a motion picture, they inspired others and heralded an important aspect of documentary film: opening the eyes to different existing worlds. In 1895, Louis Lumière, son of a French painter-photographer, launched his 'cinématographe' with one-reel films that lasted no more than one minute. The portable, hand-cranked instrument was the ideal apparatus for recording everyday life and this is what Lumière's operators did throughout the world: document life as it unfolds and project these 'panoramas' (the name used in the Lumière catalogue) for enthusiastic but puzzled audience (Barnouw, 1993).

Two years later, the Lumières started selling the equipment and new enterprises throughout the world took over the production of films, calling them "documentaires, actualités, topicals, interest films, educationals, expedition films, travel films or travelogues" (Barnouw, 1993, p. 19). As the equipment improved, longer reels made films last up to ten minutes.

For the next ten years, non-fiction films dominated the scene, but then, the audience's interest shifted to fiction films artistically edited by pioneers like Georges Méliès. Not only did its lack of innovation cause the decline of documentary films. The filmmakers had become mere agents of PR campaigns by kings', presidents' and the military, because

coronations and speeches were predictable events that proved easy to film and as a matter of promotion, filmmakers asked official leaders for sponsorship (Barnouw, 1993; Ellis, 2012). Moreover, (parts of) documentary films like military battles, hunting scenes or volcanic eruptions were often reconstituted because of practical and technical problems. Even though the public was accustomed to news pictures having an uncertain link to the event and the meaning of reconstruction was no-one's concern, documentary film suffered from the perceived idea of fakeness. The rise of fiction film downgraded it further and by 1907, it was reduced to a newsreel.

Nevertheless, interest in distant areas and the ambitions to extend documentary approaches to more exotic settings kept the documentary film alive. Explorer-documentarists like the British Herbert Ponting who filmed Captain Scott's Terra Nova expedition to Antarctica (1911) breathed new life into the genre. Yet, to be commercially viable, the films needed to be more than views of remote locations.

This is why Robert Flaherty, while documenting the life of the Canadian Inuit in *Nanook of the North* (released in the United States in 1922), cast his characters and asked them to act like their forebears. Lasting 79 minutes and telling a story through footage and editing in a way that was felt as natural, *Nanook of the North* is considered one of the first significant documentaries and a milestone in the history of the genre (Barnouw, 1993). An interesting detail is that it was funded by a private patron, Revillon Frères, a French fur company, and not by the money of Hollywood. Flaherty was known for his artistic integrity, unseduceable by the commercial film-industry and an excellent advertisement of himself (Ruby, 2000). Today, *Nanook of the North* would be labelled as a docudrama[1] because of its recreation, but at the time it was praised for its authenticity, for the definition of non-fiction was not subject to any criteria to be perceived as 'real' (Ellis, 2012). This issue, the question of defining 'documentary', will be addressed in Section 1.1.2. below.

One cannot write about documentary without mentioning the Scot John Grierson, founder and leader of the influential British documentary movement. Upon graduating in moral philosophy, Grierson moved to the United States in 1924 for research in social sciences. He was fascinated by the American melting pot and the role popular press and film could play in the citizens' education and, in this way, involve them in the country's decision-making process. When he returned to England a few years later, Grierson founded the Empire Marketing Board Film Unit, convincing the Financial Secretary of the Treasury, leading authority in the herring industry, to sponsor the Unit by making his own film about herring fisheries called *Drifters* (1929). The EMB Film Unit grew rapidly

[1] A film/movie, usually made for television, in which real events are shown in the form of a story. Retrieved from http://www.oxfordlearnersdictionaries.com/definition/english/docudrama

under Grierson's supervision, creating brilliant films, which raised the profile of the British documentary movement across the world (Ellis, 1968; Barnouw, 1993).

It was during his earlier stay in the United States, however, that Grierson coined the term *documentary*. He first used it as an adjective in a review about Robert Flaherty's film *Moana* (1926), writing about its "documentary value" as "a visual account of events in the daily life of a Polynesian youth and his family" (Chanan, 2007; Ellis, 1968, p. 20). Within a few years, the word turned into a noun still used today, although Corner (2002) observes that it is still safer to use the word as an adjective than as a noun. To ask whether or not a film is 'a documentary' emphasizes definitional criteria, whereas to ask if a film is a 'documentary project' highlights the practice rather than the object. In other words, to give an extensive description of how documentaries can be made is easier than to provide a sound definition of the term 'documentary'.

Brian Winston (1995) has explained the derivation of the word. The Latin verb 'docere' (to teach) and the noun 'documentum' (a lesson) have given origin to the English adjective 'documentary', which entered the language in the early 19th century. The modern meaning of the source word *document* (something written which furnishes evidence or information) dates from less than one century before as a consequence of the growth of legal rights grounded in contracts rather than arising from status. The term includes particularities as 'affidavit, charter, memorandum, brief, writ, note and letter', binding what is written to evidence before the law in both pre-modern and modern periods. It is in this context - a document and thus evidence – that the technology of photography and later cinematography was placed. An image (still as a photography or moving as a film) carries with it the connotation of evidence, of 'truth', of 'non-fiction'. It is the source of ideological power of documentary. Or to quote the famous Russian documentarist Dziga Vertov (1896-1954) writing about the film product:

> Each item of each factor is a separate little document, the documents have been joined with one another so that, on the one hand, the film would consist only of those linkages between signifying pieces that coincide with the visual linkages and so that, on the other hand, these linkages would not require intertitles; the final sum of all these linkages represents, therefore, an organic whole (Vertov in Michelson, 1984, p. 84).

1.1.2 Definitions

However, Grierson wanted a documentary to be more than a mere mechanical claim on the real arising from an apparatus (Winston, 1995). In a famous quote, he defines a documentary as "the creative treatment of actuality" (Rotha in Winston, 1995, p. 11), offering in an aesthetic shaping a capacity for revelation rather than merely mechanistic observation or reflection of the real (Chapman, 2009). This, however, creates a

contradiction, for one can ask how much actuality is left after a creative treatment? Moreover, as Spence and Navarro (2011) point out, all representations of actuality must choose which aspects to include or to leave out so every representation is a transformation. To represent 'the' actuality in a documentary is impossible, it will inevitably be a selective view, that of the filmmaker. Bruzzi (2000) suggests correctly that it might be better to think about documentaries as a negotiation between filmmaker and reality, between the real event and its representation, distinct but interactive.

Despite the ambiguity concerning the actuality factor in documentaries, these films are always associated with realism, which brings us to the broader issue about the boundaries between fact or non-fiction (realism) and fiction.

In his ground-breaking work *Representing Reality*, Bill Nichols (1991) devotes considerable attention to analysing the differences between fiction and non-fiction. A documentary offers information[2] about a shared, historical construct. While fiction is oriented towards 'a' world, non-fiction provides a view on 'the' world. Although documentaries may share characteristics of fiction, the filmmaker, the text and the viewer distinguish non-fiction from fiction. The filmmaker's control over what is being filmed and of the representation of lives extending well beyond the film, the text's structure, as well as the viewer's reaction and expectation are important issues, in which a documentary is "a fiction unlike any other" (1991, p. 109). Being documented texts in their own right, documentaries, like fiction, implicate characteristics of a constructed, formal, ideologically inflected status. However, they address the world in which we live rather than a world in which we imagine to live. The viewer of non-fiction is asked to consider the film as a representation, or a proposition, aiming at the historical world directly, rather than an imitation of it. Nichols aligns fiction with "likeness", while non-fiction produces representations aligned with "stickiness", images attached to referent (1991, p. 109).

A different angle is proposed by Chapman when she argues that a key factor in non-fiction is "the existence of a pre-filmic reality" (2009, p. 20) or, as Ward (2005) puts it, the events or persons depicted exist(ed) in the real world that is being documented. The purpose of replicating this reality is evidence as well as entertainment, but it is the nature of filmic representation that causes trouble in the debate between theorists of documentaries, for the attempt to achieve realism in artistic creation can be an issue of both fiction and non-fiction films. The fidelity of an image to what it aspires to represent can be questioned in a Hollywood movie as well as in the accidental record of a crime. Non-fiction, too, can be reconstructed. What makes the difference is the belief that a documentary can access the real. Yet, the nature of the moving image itself – constructed

[2] Section 1.2.2 elaborates on the informative function referring to the textual function of Werlich (1975).

from individual still-frames – is, somehow, at the heart of the recreation of reality and, as a consequence, of the verisimilitude problems (Chapman, 2009).

Plantinga (1997) argues that one cannot consider everything that manipulates its materials as fiction, for then, all films would become fiction films. The distinction between fiction and non-fiction must therefore be searched for in properties other than merely intrinsically textual ones. It is the extrinsic context of production, distribution and reception that sheds light upon this issue. The stance taken by the filmmaker toward the world projected in a fiction film is a fictive stance. The spectator is invited to consider the states of affairs it presents and not to verify its existence. The stance taken toward non-fiction films is an assertive stance, the states of affairs represented are asserted to occur and the viewer can evaluate their truthfulness.

In this context, it is important to remind ourselves, as Carroll (1996) points out, that producers index their films, identifying them publicly as either fiction or non-fiction, and indexing tells us the kind of expectations to bring to a film. Spectators respond to a film according to the indices. In the case of non-fiction, they will make an appeal to objective standards of evidence and argument to gauge the representation. For each of the knowledge claims made in a non-fiction film, there could be evidence, whereas for fiction films, the issue of knowledge claim or evidence is dropped (1996).

Yet, indexing is a social phenomenon, determined by both audiences and producers, which may change in time since social conventions change. Moreover, some films purposefully combine fictive and assertive stances, creating fuzzy boundaries between fiction and non-fiction, yet, it is no less a valid distinction for clear cases of non-fiction and fiction (Plantinga, 1997).

Ward (2005) agrees with Plantinga that purpose and context make a clearer distinction than form or style (the intrinsically textual properties). Staging, on the one hand, and capturing profilmic[3] material, on the other, are both viable options in a documentary that should not constitute an element of distinction. What makes a film a documentary is the interaction between text, context, producer and spectator and is socially negotiated by the audiences. As Nichols (2001) explains, the multiple agents producing documentaries define the genre: "documentaries are what the organizations and institutions that produce them make. [...] The context provides the cue." (p. 22). Documentary is a protean institution, consisting of texts, a community of practitioners – the distinct agents that creating/using documentaries - and conventional practices, which are subject to historical change.

In line with Ward, Plantinga and Nichols, Chanan (2007) states that nowadays, there are too many different styles and techniques to establish clear criteria for the difference between fiction and non-fiction. He suggests a very useful view on defining documentary,

[3] About a world "that happens, irrespective of whether or not the camera is present" (Ward, 2005, p. 8).

based on Wittgenstein's theory of forms of life, by comparing it to a family tree. Different members share their features, however none of these features define by themselves the characteristic of family resemblance. It is the common genealogy that makes the members belonging to the same family. Documentary films can be quite different from one another, just like members of a family, yet they constitute an extended family with different branches or clusters of conventions. The genealogy is the common component, and the main branches represent particular traditions that serve as models. Those who follow the model, however, might not resemble one another. Features from different branches are shared and mixed and create new paradigms. If one considers documentaries in terms of genealogy, then deciding which films to label as 'documentary' becomes less confusing.

1.1.3 Classification

The variety of documentary practices, used to represent situations and characters and the fuzzy boundaries of its definitions, turn classification of documentaries into a tedious task.

1.1.3.1 Representation modes throughout the ages

The most influential and seemingly sole attempt to classify the basic ways of organising documentary texts was elaborated by Nichols (1991). He distinguished four modes of representation: the expository, the observational, the interactive and the reflexive mode, to which he subsequently added the poetic and the performative mode (2001). The modes roughly follow a chronological order. While their characteristics are to be considered dominant, influence of other modes may be included. None of the modes cancel out the previous one, they overlap and interact (Nichols, 1994).

The poetic mode emphasizes mood rather than conveying information through rhetoric. Social actors, such as psychologically complex characters, temporal rhythms and spatial juxtapositions, are underdeveloped in favour of alternative forms of knowledge. Reality is represented by a series of fragments, subjective impressions, incoherent acts and loose associations, creating in this way a sense of honesty.

The expository mode constitutes a rhetorical frame, addressing the audience directly with a voice-of-god commentary (the speaker is heard but never seen) or voice-of-authority commentary (the speaker is heard and seen) to offer information and to represent the argument of the film. Interviews and social actors are, if anything, subordinate to this argument and the commentary voice. Images only serve a supporting role.

The observational mode stresses the non-interference of the filmmaker. The camera observes spontaneously lived experience, social actors engage with one another,

disregarding the presence of the filmmaker. In order to honour this spirit, no commentary voice, or interviews, are introduced during the footage or in post-production. As the filmmaker is no more than a 'fly on the wall', the scenes encourage the viewer to adopt an active position in understanding the significance of what is shown.

The interactive mode, also called participatory mode, emphasizes the encounter of the filmmaker and the subject. The filmmaker goes into the field and tells or represents an experience, becoming a social actor, responding on the spot. He serves as a researcher, his voice emerges from direct, personal involvement in the events, in some cases combined with interviews. The audience is a witness to the dialogue between the filmmaker and the subject. Images bear out or demonstrate the validity or doubtfulness of what witnesses state.

The reflexive mode focusses on the negotiation between the filmmaker and the viewer as a social actor. The filmmaker addresses the viewer directly or relies on interviews, pretending in some cases they are spontaneous but showing later they are staged. Reflexive documentaries invite us to heighten our awareness of the problems of representing others and to reflect on the film process and the editing techniques. These documentaries may have a strong social or political impact, enhancing the consciousness of the gap between knowledge of what exists and desire to know what might exist.

The performative mode stresses the subjective and affective dimensions of our knowledge of the world besides factual information. It calls for emotional response, leaving analysis and judgement to the viewer. Commentary voice and interviews serve to evoke a subjective experience (Nichols, 1991; 2001).

Bruzzi (2000) questions Nichols' premise that his classification evolves in a chronological order and Ward (2005) agrees that this can be misleading. Nevertheless, both scholars argue that the categories are very useful if one accepts that documentary practices use a range of modes that relate to each other dialectically.

One example to illustrate this juxtaposition of modes is the documentary *Lift*[4] (Marc Isaacs, 2001). Isaacs spent one month in a lift in a tower block in London, holding a camera directed towards the lift doors, filming whoever entered the lift. At first, he remained silent, but when the passengers (the inhabitants of the block) gradually got used to the camera, Isaacs started asking questions. The documentary develops in a more interview-based film, offering an insight into the world of the passengers. What counts in this film is the changing social fabric and people's attitude towards it. The first mode to be distinguished is the observational mode as, initially, Isaacs did not engage with the people, he just recorded. When he started asking questions, the interactive mode takes over. The reflexive mode can also be discussed here as the film reveals to some extent

[4]Retrieved from https://www.youtube.com/watch?v=FJNAvyLCTik

that it is a construction and in the meantime, the viewer is invited to think about the social implications of the construction.

1.1.3.2 New trends and the increasing importance of documentary festivals

The digital revolution of the past 20 years and the fast evolving 'screen world' tailored to demand, distribution, budget and television competition ratings of films have encouraged film-makers to expand the traditional boundaries of documentary films.

The popular docusoaps[5] and reality shows[6] appear to have taken over much air time on terrestrial television that was formerly reserved for traditional documentaries (informative or educational issues), which are now diverted to themed channels. Digital cameras and desktop editing offer independent documentarists the opportunity to make low-cost films. Their distribution is no longer confined to traditional commercial operations, but reaches the cinemas thanks to the demand of some of the cinema-going public. In addition, the World Wide Web hosts and encourages a range of experiments of digital documentaries and interactive web documentaries, created for the web, accessed through the web and employing both web-design and documentary modes of representation (Šukaytitè, 2012).

As a result of this evolution, an open model is needed to determine what a documentary is and what is not. Within this framework, Chanan's 'documentary tree' is interesting, because it allows innovative film styles and techniques to be added to the documentary 'family'.

In order to illustrate the (r)evolution in this field, I have consulted an up-to-date news source for documentary film i.e. the documentary festivals, focussing on their dates of inception and their attention for new trends. Table 1 below lists the main international documentary festivals in the world with their dates of inception. Sixteen out of 26 festivals were staged after 1997, 13 of which after the year 2000 (see Table 2), which demonstrates the public's booming interest in the genre over the last 20 years.

Table 1 International documentary festivals: name, location and date of inception

Docville Leuven[7] (Belgium)	2005
Milleniumfestival Brussels[8] (Belgium)	2009

[5] An entertaining TV programme about the lives of real people, especially people who live in the same place or do the same job. Retrieved from http://dictionary.cambridge.org/dictionary/english/docusoap.

[6] A television programme in which ordinary people are continuously filmed, designed to be entertaining rather than informative. Retrieved from https://en.oxforddictionaries.com/definition/reality_show.

[7] Retrieved April 9,2015, from http://www.docville.be/

[8] Retrieved April 9,2015, from http://www.festivalmillenium.org/

IDFA Amsterdam[9] (The Netherlands)	1988
CPH:DOX[10] Copenhagen (Denmark)	2003
Cinéma du reel Paris[11] (France)	1979
Sunny Side of the Doc[12]La Rochelle (France + Japan & Mexico)	1989
LIDF London[13] (United Kingdom)	2007
Sheffield Doc/Fest[14] (United Kingdom)	1993
IFI Documentary Festival[15] Dublin (Ireland)	2003
Documenta Madrid[16] (Spain)	2004
DocLisboa[17] (Portugal)	2004
Festival dei Popoli[18] Florence (Italy)	1959
Thessaloniki Documentary Festival[19] (Greece)	1999
Dok Leipzig[20] (Germany)	1955
DOK. Fest[21] Munich (Germany)	1985
Aljazeera International Documentary Festival[22] Doha (Qatar)	2005
DocAviv[23] Tel Aviv (Israel)	1998
SilverDocs/AFI Discovery Channel festival[24] Washington DC (United States)	2003
True/False Film Festival[25] Columbia, Missouri (United States)	2003
Hot Docs[26] Toronto (Canada)	1993
It's all true[27] Rio de Janeiro and São Paulo (Brazil)	1996
DocsDF[28] Mexico City (Mexico)	2006

[9] Retrieved April 9,2015, from http://www.idfa.nl/nl.aspx
[10] Retrieved April 9,2015, from http://cphdox.dk/en
[11] Retrieved April 9,2015, from http://www.cinemadureel.org/en
[12] Retrieved April 9, 2017 from http://www.sunnysideofthedoc.com/fr/
[13] Retrieved April 9,2015, from http://www.lidf.co.uk/
[14] Retrieved April 9,2015, from https://sheffdocfest.com/
[15] Retrieved April 9, 2017, from http://ifi.ie/docfest
[16] Retrieved April 9,2015, from http://www.documentamadrid.com/
[17] Retrieved April 9,2015, from http://www.doclisboa.org/2015/en/
[18] Retrieved April 9,2015, from http://www.festivaldeipopoli.org/en/
[19] Retrieved April 9,2015, from http://tdf.filmfestival.gr/default.aspx?lang=en-US&page=1244
[20] Retrieved April 9,2015, from http://www.dok-leipzig.de/home/?lang=en&
[21] Retrieved April 9,2015, from https://www.dokfest-muenchen.de/
[22] Retrieved April 9,2015, from http://www.aljazeera.net/festivalenglish
[23] Retrieved April 9,2015, from http://www.docaviv.co.il/org-en/
[24] Retrieved April 9,2015, from http://afi.com/afidocs/default.aspx
[25] Retrieved April 9, 2017 from https://truefalse.org/
[26] Retrieved April 9,2015, from http://www.hotdocs.ca/
[27] Retrieved April 9,2015, from http://www.itsalltrue.com.br/en/home/
[28] Retrieved April 9,2015, from http://docsdf.org/en-el-hoyo/?lang=en

YIDFF[29] Yamagata (Japan)	1989
GZDOC[30] Guangzhou (China)	2003
TDIF Taipei[31] (Taiwan)	1998
Documentary Edge[32] (Auckland and Wellington) New Zealand	2005

Table 2 Documentary festivals in order of inception date

< 1960: Florence / Leipzig
1960 – 1980: Paris
1980 – 1990: Amsterdam / Munich / Yamagata / La Rochelle
1990 – 2000: Sheffield / Thessaloniki / Tel Aviv / Toronto / Rio de Janeiro and São Paulo / Taipei
2000 – 2010: Leuven / Brussels / Copenhagen / London / Dublin / Madrid / Lisbon / Doha / Washington DC / Columbia / Mexico City / Auckland and Wellington / Guangzhou

Some of these festivals dedicate a special section to new trends. For example, Docville Leuven states on its website that the documentary sector reinvents itself continuously, exploring the traditional boundaries. The section 'Outside the dox'[33] welcomes documentaries that differ from the mainstream in form, content or narration. For example *Love is all: 100 Years of Love and Courtship* (Kim Longinotto, UK and US 2014) is a collage of archive images of love scenes on film throughout the 20th century, exploring the depiction of love and courtship. The section 'Thrilling docs'[25] presents another new trend i.e. the use of elements of narrative cinema to add drama to documentary (elements of thrillers, horror films, detectives, heist films). *The arms drop* (Andreas Koefoed, Denmark, India, UK, Sweden 2014) is a docuthriller about the 4-ton arms dropped over India in a December night in 1995. It tells the nerve-racking story of two men, each with their own hidden agendas, who gamble their lives on a joint mission, and the political, personal and diplomatic consequences 20 years later with reconstructions and scenes resembling a feature film.

IDFA Amsterdam organises as from 2006 a section called 'Paradocs'[34] for the 'periphery' of the documentary to showcase what is going on beyond the frame of traditional documentary filmmaking, on the boundaries between film and art, truth and fiction, narrative and design.

[29] Retrieved April 9,2015, from http://www.yidff.jp/home-e.html

[30] Retrieved April 9,2015, from http://www.gzdoc.com/index!home

[31] Retrieved April 9,2015, from http://www.tidf.org.tw/en

[32] Retrieved April 9,2015, from http://documentaryedge.org.nz/2015-home/

[33] Retrieved April 10,2015, from http://www.docville.be/programma/

[34] Retrieved April 10,2015, from http://www.idfa.nl/nl/festival/programmaonderdelen/paradocs.aspx

DocLisboa's section 'New vision'[35] was created in 2007 to extend the scope of DocLisboa with different approaches of reality and its representations, defying the usual categories, formats, and length, into frontiers between fiction and documentarism, and different trends as filmed diaries and autobiographies, or works on archive footage.

TDIF Taiwan has a section 'Stranger than documentary'[36] for short films that challenge the conventional definition and format of documentary. 'Strange' implies a subversive and ground-breaking nature. Moreover, the section questions the concept of 'reconstructed reality' and seeks to stimulate people's imagination.

Sheffield Doc/Fest has programmed a crossover summit since 2009 which explores new approaches to commissioning content aimed at maximising audience engagement, the so-called 'interactive documentaries'[37]. A dedicated website for interactive documentaries (i-docs) has been set up as one sub-section of the i-Docs project which is a research strand within the Digital Cultures Research Centre at UWE Bristol. For these scholars, an i-doc is 'any project that starts with an intention to document the 'real' and that does so by using digital interactive technology'[38]. The team also organises a bi-annual symposium for i-docs.

Given all these new trends in both content and technique, a classification based on representation modes which Nichols proposed is considered too rigid, since documentary films usually combine different modes – as Nichols says himself (1994) – and since the search for innovation is intrinsic to documentary filmmaking.

Since proposing a new classification that covers 'the' documentary genre is beyond the scope of this research, the focus to address the question of classification will be on the Flemish public broadcasting company VRT, who provided the corpus for this research, and the needs of the thesis under study (terminology in documentaries).

VRT has two documentary channels: 'Eén' and 'Canvas'. Eén offers programmes aimed at the general population and wants to be the mirror of Flanders[39], while Canvas homes in on people in search of information, analysis, and self-awareness, focussing on innovation and creativity[40]. Documentaries broadcast by Eén tell a real, recognizable story about people who actively participate in the film. The camera observes and an off-screen commentary voice guides the audience through the film. For example, *Het leven zoals het is* ('Life as it is', own translation) shows real-life scenes in Brussels airport, with the Antwerp police or at a children's hospital. Not only Flemish, but also international

[35] Retrieved April 10,2015, from http://doclisboa.org/2014/en/edicao-actual/seccoes/riscos/

[36] Retrieved April 10,2015, from http://www.tidf.org.tw/en/category/shows/15

[37] Retrieved April 10,2015, from https://sheffdocfest.com/articles/102-interactive-at-sheffield?tag=sessions

[38] Retrieved April 10,2015, from http://i-docs.org/about-idocs/

[39] Retrieved April 27, 2015, from http://www.vrt.be/en/een

[40] Retrieved April 27, 2015, from http://www.vrt.be/en/canvas

docusoaps about animals like *Zoo Australia* and biographic films about celebrities like the Spice Girls follow the same pattern of observation and commentary. In lifestyle series like *Restaurant Inspector* (Peire 2012) or in travel documentaries, an authority voice (e.g. Jamie Oliver, Joanna Lumley) provides the commentary.

Documentaries broadcast by Canvas, however, offers added value in terms of both content and form. They are informative, instructive while being objective and discerning at the same time. Recurring themes are history, science, wildlife, sports, arts, music, travel, current affairs and director's documentaries. Historical documentaries cover 20th-century themes, science documentaries discuss topical subjects like climate issues and wildlife documentaries bring epic, narrative films like *Planet Earth* (2006). An example of a sports documentary is *Senna* (about the Brazilian Formula One driver, Kapadia 2010) and *The Private Life of a Masterpiece* (about great works of art, BBC 2001-2010) is an example of an arts documentary. Music documentaries (about both modern and classical music) focus on anecdotes and characters, while travel documentaries explore special human interest issues like Louis Theroux investigating peculiar social environments (e.g. sex offenders in Los Angeles). Current-affairs documentaries address topical subjects concerning news issues (e.g. Rudi Vranckx' reports in conflict areas). Director's documentaries, finally, offer a personal view of the filmmaker himself (e.g. Werner Herzog, Michael Moore)[41].

For the purpose of this dissertation (the translation of terminology in documentaries), however, a pragmatic, subject-based classification has been gleaned out of the complete list of documentaries broadcast by VRT (both Eén and Canvas) between 2005 and 2012. The following categories can be distinguished: arts, current affairs, director's documentaries, history, lifestyle, music, wildlife, science, society, sports and travel. The hypothesis in the thesis under study is that, while all these subjects can contain terminology, the specificity and number of terms – important features for the performance of automatic terminology extraction systems - will be tailored to the target audience (the Canvas documentaries will use a more specialised language than those broadcast by Eén) and the subject (wildlife or science documentaries will contain more terminology than lifestyle or travel films). In Chapter 2, these features are analysed in order to select the corpus for the translation experiments.

[41] Tom Bleyaert, VRT Programme Acquisition Department, personal conversation,April 24, 2015.

1.1.4 The verbal mode

As the core issue of the present thesis – terminology – is a lexical feature, spoken through the commentary voice, the following section deals more extensively with the verbal mode.

1.1.4.1 Definition and functions

In the above definitions and classifications, the different role of verbal modes in documentary film is regularly raised, but never discussed in great detail. The verbal mode recurs in the different definitions and classifications in narrative forms, i.e. voice-of-god commentary and voice-of-authority commentary (the terms are explained below in 1.1.4.2.), or in interactive forms, i.e. interviews or the filmed conversations of interactants - to name but the most common ones. As each of these verbal modes pose specific challenges for audiovisual translation (see Section 1.2), they are briefly discussed below.

The verbal mode as a narration technique is here limited to what is verbally said. While it would take us too far to analyse the voice of documentary spoken through the selection and arrangement of sound and image, broadly speaking, the verbal mode interacts with other narration systems like sound and image, but these other systems are not taken into consideration as the focus of this research is terminology. For the same reason, the translation experiments presented in Chapter 4 have been carried out without the use of the video. The source text of the experiments consisted of unrelated clips, the terminology was clear from the verbal context and there was no need to match timing or style with the images and the intonation of the voice talent for the purpose of this study.

For the discussion of the verbal mode, I turn to Cattrysse, Kozloff, Nichols and Remael, because script-writing manuals pay little attention to documentaries; Cattrysse (1995) is an exception. In the literature about filmmaking techniques, scholars have only minimally addressed the verbal mode. In *Cross-Cultural Filmmaking. A Handbook for Making Documentary and Ethnographic Films and Video's* (1997), Barbash and Taylor dedicate a section to interviews and voice-over narration from the filmic point of view. They also discuss translation, limiting the focus to pros and cons of subtitling and dubbing and their technical issues. The verbal mode from the lexical point of view is not addressed. Nichols (1991) mentions it briefly, Remael (1999 and 2000) conducted early case studies and doctoral research about film dialogue, but Kozloff's *Invisible Storytellers* (1988) and *Overhearing film dialogue* (2000) are still the most important references.

In her extensive and precise analysis of voice-over narration in fiction films, Kozloff (1988) makes a first connection between voice-over narration and documentaries explaining that 'voice-over narration provides the perfect means of conveying all discursive and expository information relevant to non-fiction matter' (1988: 28). She explains that voice-over narration is a voice the viewer hears but does not see in the act

of recounting a narrative. She distinguishes first-person and third-person narrators. "Accordingly", she states, "voice-over narration can be formally defined as oral statements, conveying any portion of a narrative, spoken by an unseen speaker, situated in a space and time other than that simultaneously being presented by the images on the screen" (1988, p. 5).

Although Kozloff relates voice-over narration to documentaries, a short consideration about this term needs to be made. Examining the modes of transfer applied to non-fiction programmes, Franco (2000) considers terminological issues in film studies. The term 'voice-over narration', she argues, has been replaced by 'voice-over commentary' when referring to documentaries, apparently because of the functional distinction between the use of the (unseen) voice in fiction versus non-fiction films. In addition, Kozloff's definition of voice-over narration is useful to put forward as a second reason to use voice-over commentary for non-fiction and voice-over narration for fiction. This definition says that voice-over narration is spoken by an unseen speaker (1988). Yet, the narrator in documentaries can be both off or on-screen. Filmmakers like Michael Moore or authoritative voices like Stephen Fry appear before the camera in order to add credibility to their argument.

An interesting point of view concerning the verbal mode in film is Kozloff's (2000) classification of dialogues into 'functions of dialogues in narrative film'. Although she discusses fiction films, some of her functions can easily be applied to non-fiction. One is 'the anchorage of the diegesis and the identities', recognizable in documentaries when the commentary voice situates the images e.g. in wildlife documentaries like *Madagascar, Island of marvels* the commentary voice in the opening scene says:

> '60 million years ago, on the shores of this tropical island, an extraordinary story began [...]'.

One minute later the presenter on-screen, David Attenborough, adds:

> 'The island was Madagascar',

identifying with these words the diegetic world. Documentaries may also need to anchor identities like interactants telling their story as a support of the filmmaker's view, talking heads explaining a concept or characters doing something. In *The earth machine - Land*, a man is seen taking lava samples while the commentary voice introduces him:

> 'Some scientists think the lava may come from a column of superheated magma from deep down within the mantle. Dario Tedesco is one of the world's leading authorities on Niyrangongo',

anchoring in this way the identity of the man on the screen.

A second function useful for documentaries is the 'communication of narrative causality' (Kozloff, 2000). Whereas in fiction, it is the characters hauling the causal chain through the film, in non-fiction it is mainly the commentary voice explaining 'why', 'how' and 'what next'. A documentary series broadcast by Eén (see also 1.1.3.2.) *Big cat diary* (1996) follows several animal stories in different locations, while the commentary voice provides the necessary information for the viewer to keep up with the story line.

'The enactment of narrative events' is the third function applicable to non-fiction. Kozloff (2000), inspired by Austin's and Searle's speech act theory, explains that narrative events can be verbal acts, for talking means "doing something (promising, informing, questioning)" (p.41). In non-fiction films the main part of the verbal act is exactly this: informing the viewer, and questioning or explaining a certain event, phenomenon, situation, personality. It is the core of documentary filmmaking.

The last function of Kozloff's classification useful for non-fiction goes beyond narrative communication into the realms of ideological persuasion: 'conveying thematic messages, authorial commentary, allegory'. It considers a statement, inviting reflection on serious issues and tends to occur in the last quarter of the film when the thematic stakes have been made clear. In documentaries, the commentary voice, as well as interviews, can be employed to convey such a message or to convince the audience about the importance of a certain case e.g. the closing comment in *Can we save planet earth?* (2006) saying:

> 'In the past we didn't understand the effect of our actions. Unknowingly we sowed the wind, and now literally we are reaping the whirlwind. But we no longer have that excuse. Now, we do recognise the consequences of our behaviour. Now, surely we must act to reform it. Individually and collectively. Nationally and internationally before we doom future generations to catastrophe.'

Tracing back these four functions to the focus of this research, namely terminology, one can expect the enactment of narrative events and the conveying of thematic messages to contain the most terminology. The former will use terminology to inform and explain and the latter will repeat this terminology if necessary to convey the message.

For the creation of the diegetic world and the communication of narrative causality (the first and the second function mentioned above), more general language will be used since these functions consider brief introductions to the frame of the story and not the core where more in-depth analysis is presented. Chapter 2 deals more extensively with the number of terms in the corpus.

1.1.4.2 Screenplay narration techniques

According to Cattrysse (1995), four basic screenwriting techniques can be distinguished, selected in function of the nature of the documentary project, the target audience, the practical circumstances and the budget. To these four techniques, a fifth was added

inspired by Nichols (1991). The terms identifying the techniques are those used by most scholars and documentarists.

- The voice-of-god commentary is a commentary voice the viewer hears without seeing the speaker. A voice talent reads the commentary in the film studio, taking care of maintaining the time link with the visuals.
- The voice-of-authority commentary is a commentary voice of the filmmaker himself or a presenter, often an authority in the field of the documentary, on or off-screen. As mentioned before (see 1.1.4.1.), the authoring presence adds credibility to the programme, which is important for documentaries, intrinsically carrying the connotation of truth. Moreover, looking into the camera and addressing the viewer directly, the speaker grabs the audience's attention. When the filmmaker/presenter is off-screen, the audience recognizes his/her voice through intonation, accent and style (Franco 1999).
- Interviews with 'talking heads', authorities in the field invited to explain a subject or give their opinion is the third verbal technique. They add information and explain a subject. Being prepared and constituting the utterance of an expert speaker, these interviews contain precise phrasing and syntax.
  Interviews with talking heads can serve both as communicator of narrative events and as a conveyer of thematic messages or authorial commentary.
- Staged scenes with dialogues form the typical verbal technique for what is called 'docu-fiction'. This technique can alternate with (one of) the former ones. Staged scenes, like spontaneous interviews, serve to add emotion and evidence for the filmmaker's argument.
  In Kozloff's terms, their main function is to communicate narrative events.
- Spontaneous interviews with common people is the fifth technique. They may contain verbal incoherence, hesitations, grammatical errors and unfinished sentences. Being spontaneous, they add emotion (Remael 1999), but they mainly serve as evidence for the filmmaker's argument (Nichols 1991).

The first two verbal techniques constitute the 'voice-over commentary' (see also 1.1.4.1.). Addressing the viewer on or off-screen with a voice that advances an argument, the voice-over commentary is the dominant verbal technique for Nichols' expository mode (see 1.1.3.1.).

As far as Kozloff's classification of dialogues is concerned, commentary fits into all four functions mentioned on p.19-20. The anchorage of the diegesis, the communication of the causal chain, of the narrative events and of the ideological message can all be done through a voice of an (un)seen speaker.

In the VRT corpus of this study, both voice-over commentary and interviews are used, with a predilection for voice-over commentary and interviews with talking heads in documentaries broadcast by Canvas and more spontaneous interviews in documentaries

broadcast by Eén. No staged scenes are shown on the Flemish television nor do they feature in the corpus of this study as docufiction is not considered in the VRT programmes. Docufiction is video work in which actual recorded events are combined with recreations or imaginary scenes in order to provide information on a specific topic[42]. This can be expected to lend itself less to the use of terminology extraction.

In his handbook for screenwriting, Cattrysse (1995) formulates some useful tips for the writing of spoken language in Dutch. Similar comments can be found in the *VRT leidraad voor commentaarvertaling* (2010)[43], which will be covered more extensively in the next section 'Insights from translation studies', where each of the verbal modes discussed here will be related to the forms of audiovisual translation usually used to render them.

In his advice to novice documentary filmmakers, Cattrysse highlights the importance of short, affirmative, active sentences in a grammatically correct language for non-fiction film. The audience has to understand what is being said at once, without too much effort because simultaneously, other information through images is provided. For the same reason, domain-specific terminology must be limited to what is strictly necessary. More detailed information regarding this topic is provided in the next section. Staged or spontaneous interviews may be (partially) prepared or impromptu, depending on the experience of the interviewed person. Long, ungrammatical sentences and unstructured narration must be avoided. Dialogues for staged scenes, representing spoken language, may differ from the strictly grammatical rules commentary has to follow.

The complexity of verbal narration, however, will be linked to the complexity of the documentary at hand and the techniques it uses to tell its story (see Nichols' classification under 1.1.3.1. and the verbal techniques above) and its aims, e.g. in terms of its socio-cultural or scientific aims. The latter will then have an impact on the type of vocabulary or terminology the film uses, which is important for the core issue the present thesis investigates.

1.2 Insights from audiovisual translation studies

The second element of the research question regards the translation of documentaries. The section below covers an analysis of the genre in all its aspects, narrowing the focus

[42] Retrieved June 15, 2017, from https://muse.jhu.edu/article/19610/summary

[43] The guide for translation of the commentary voice in documentaries, edited by the Flemish broadcaster VRT.

then to terminology and presenting an overview of existing projects concerning the automation of audiovisual translation.

1.2.1 The specific challenges of translating documentaries

Documentary is a protean genre, complex both in content and form, as shown in the previous chapter. This complexity not only constitutes a challenge for documentary theorists trying to define the genre, it is also an issue in audiovisual translation (AVT) research. Several scholars have highlighted either the specificity of translating this genre or the versatility required on the part of the translator, not to mention the different challenges posed by the different audiovisual translation modes used for documentaries.

Franco has brought the translation of documentaries to the fore. Through case studies, she demonstrated that documentary translation needs to be seen as a specific practice and that translational choices for subtitling, voice-over or off-screen dubbing[44] may influence the interpretation of the documentary reality. She has also studied terminological and conceptual issues, focusing on voice-over (1999, 2000, 2001a, 2001b).

In line with Franco, Espasa (2004) has argued that documentary translation constitutes a full audiovisual translation product, adding as specific challenges the heterogeneous audience and the domain-specific terminology, regardless of the translation mode. A corresponding view is proposed by Matamala who has described the main challenges in the translation of documentaries as being proper nouns, unintelligible excerpts, incorrect transcriptions and terminological problems (2009a, 2009b, 2010).

In their overview of voice-over translation, Franco, Matamala and Orero (2010) turned some of their attention to the translation of non-fictional products and to the translation process of voice-over and off-screen dubbing for non-fiction. Voice-over is also the core issue in Orero's analysis of the voice and the accent of the speaker as one of the features to make audiovisual media a construct of reality (2006).

Kaufman has carried out two case studies regarding the subtitling of documentaries. She highlighted the importance of a complete and accurate source text so the translator can provide a complete and accurate target text, which is essential for keeping the illusion of authenticity in non-fiction (2008). In another study, she discussed the consequences of the French language policy (standardizing the characters' idiolects and sociolects) when applied to the subtitles (2004).

In line with the latter study, Remael (2007) has analysed the homogenizing effect of the Flemish television language guidelines for subtitling and off-screen dubbing.

[44] Term proposed by Franco et al. (2010) which indicates the transfer that takes place from original commentaries into translated ones, when on-screen speakers and the original audio tracks are absent and consequently also any evidence of lip synchronisation.

Moreover, she argued that Flemish translators and/or journalists refract the documentary in order to tailor it to the house style. Taylor (2002) has addressed subtitling as well, suggesting the use of computer-based multimodal transcriptions in order to formulate translation strategies, especially where nature documentaries for television are concerned.

Authors writing about AVT translation strategies always refer to existing strategies that are then adapted to the specific text genre under study. Even recent studies like Pederson (2011) and Ranzato (2012), which both deal with the same topic, the translation of cultural references, propose strategies that overlap and overlap with existing ones for cultural references by other authors. In addition, translation strategies commonly used in other domains e.g. technical translation, are specific to that domain whereas off-screen dubbing is a mixed genre consisting of both creative elements and domain-specific elements as mentioned in Section 1.1.1. Moreover, they aim at a mixed audience, as was indicated above.

A free translation, that is easy to understand is one of the main features the VRT guidelines for off-screen dubbing promote (discussed into detail in Section 2.4). Repetition for instance is to be avoided because it is thought to be disturbing for the audience, so translators have to find synonyms or other creative solutions. This means that they will try to use the correct term but will also be interested in finding synonyms that are 'equally' correct or almost. This is important for the terminology lists they might want to use: they have to be slightly fuzzier than terminology lists for technical translation 'pure and simple'.

Nevertheless, investigating translation strategies commonly used in technical translations might be interesting for future research, for instance, regarding the translation of corporate videos, focussed on a target audience of experts.

There are different reasons why translating documentaries is a very specific audiovisual practice. Translators of written texts are usually specialists in just a few fields, whereas documentary translators are generally not specialised in a given field or topic, but in a specific mode, i.e. 'audiovisual' translation. As such, they need "minimum knowledge of a maximum of topics" (Mir in Espasa, 2004, p. 190).

Moreover, the mode of discourse to be translated, being the oral rendering of a previously written text, can include various registers - more formal for the narrator and more spontaneous for interviewees - and consequently add diversity to the translation, as is reflected in the translation guidelines (Gregory & Carroll, 1978; Espasa, 2004).

Besides the audiovisual mode and the mode of discourse, documentaries can use various translation modes, such as voice-over translation[45], lip synch dubbing, off-screen dubbing and subtitling, each of which requires specific knowledge. Depending on their working country, audiovisual translators, most of whom also translate documentaries, are specialised in one or more translation modes. In what are traditionally considered to be dubbing countries (France, Italy and Germany, to name but a few), documentaries feature mainly voice-over translation, lip synch dubbing and off-screen dubbing (Franco et al., 2010), whereas in subtitle countries like the Scandinavian countries, the Netherlands and Belgium, subtitling in combination with off-screen dubbing is used.

Until now, scholars studying the translation of documentaries have discussed mainly voice-over translation or subtitling. Little research has been done into the specific features of off-screen dubbing, the subject addressed in this dissertation. The following studies on off-screen dubbing appear to constitute the main body of research for this audiovisual mode. Remael (2007) analyses the translation shifts of both subtitles and narration[46], due to language and/or ideology policies. Franco et al. (2010) discuss briefly that for off-screen dubbing translators do have to take into account synchronicity constraints concerning the visuals and the timing, although the feeling exists that synchrony has no importance because the viewer does not hear the original voice. Additionally, the language register of the source text must be respected and the translation must be for a speaker who will read it aloud. Moreover, they mention two of the specific features off-screen dubbing shares with voice-over: terminology and a good comprehension of the source text (2010).

The VRT guidelines for off-screen dubbing also address synchrony between the spoken text and the visuals, including some linguistic suggestions to avoid losing synchrony (Osstyn, 2010).

Off-screen dubbing for documentaries is examined by the ALST-project (Matamala et al., 2012) as well. This study investigates the application of machine translation and post-editing of off-screen dubbing for wildlife documentaries. As a part of the ALST-project, Ortiz-Boix (2016) analyses the translations produced by MT engines and mentions terminology as one of the challenges. She also carried out an experiment with Master's students in translation who translated and post-edited an excerpt of off-screen dubbing for wildlife documentaries. The results showed generally that post-editing requires less temporal, technical and cognitive effort than translating (Ortiz-Boix & Matamala, 2016).

Besides the three challenges for translators of documentaries (the audiovisual mode, the mode of discourse and the different translation modes), there is a fourth challenge

[45] "Voice-over translation[...] is the revoicing of a text in another language, or a translating voice superimposed on a translated voice, usually starting a few seconds after the original [...]" (Franco et al. 2010, p. 43).

[46] The translation of the off-screen voices (= off-screen dubbing).

concerning the variety of textual functions of the verbal mode. According to the classification by Rosa Agost for the translation of audiovisual genres (1999), documentaries are considered to be an informative genre with specific textual functions, including narrative, descriptive, persuasive and expository functions.

These functions go back to Werlich's linguistic analysis 'Typologie der Texte' (1975) where each function corresponds to a sentence type according to its specific lexical involvement. The narrative function corresponds to an action-recording sentence (involving time indicators), the descriptive function to a phenomenon-registering sentence (involving space indicators), the persuasive function to a quality-attributing sentence (involving negation and/or contradiction indicators) and the expository function to a phenomenon-identifying or a phenomenon-linking sentence (involving classification indicators).

Agost uses the textual functions for the classification of the different audiovisual genres, whereas the focus of the present dissertation is the translation of documentaries. Considering that documentaries contain all four of Werlich's textual functions (see also 1.1.4 for the verbal mode in documentaries), their translation will require attention in terms of the specific indicators of each function.

Another classification confirming the informative and the persuasive functions of documentaries can be traced back to Gommlich's text analysis. Espasa (2004) explains that this translation-oriented classification of scientific and technical texts proposed by Gommlich (1993) is also applicable to documentaries. Gommlich distinguishes four categories that influence the translation strategy:

- Transfactual texts I: They have an informative function and address an expert audience e.g. business videos for specialists in a specific field
- Transfactual texts II: They have an informative function and address a non-expert audience e.g. educational documentaries about World War II
- Transbehavioural texts I: They have a persuasive non-binding function, attempting to influence the behaviour of the target audience e.g. documentaries about global warming
- Transbehavioural texts II: They have a binding persuasive function (laws and patents) e.g. legal restrictions mentioned in the closing credits at the end of a film.

Most documentaries can be classified under 'transfactual texts II' and 'transbehavioural texts I'. The latter is what Chanan (2007) calls new-wave documentaries, where the traditional stance of impersonality (as in 'transfactual texts II') is replaced by a more persuasive stance of the filmmaker him/herself to express a point of view. In Nichols' classification (1991), the 'transbehavioural texts I' correspond to the interactive and the reflexive mode of documentary film- making.

In summary, the translation of documentaries requires an all-round knowledge of text types and functions, language registers and translation modes since, each of the four

mentioned features presents its own challenge: the specific audiovisual mode proposed by Franco (1999, 2000) and Espasa (2004), the mode of discourse containing diverse registers as tackled by Espasa (2004), the various translation modes possible in a documentary (voice-over translation, lip-synch dubbing, subtitles and off-screen dubbing) and the textual functions introduced by Werlich (1975) and applied to audiovisual texts by Agost (1999), as well as those introduced by Gommlich (1993) and applied to documentaries by Espasa (2004).

1.2.2 Verbal narration and its lexical features

Narrowing the focus to the core issue of the thesis under study – terminology in documentaries and the possible support automatic terminology extraction systems can offer for the translation of documentaries – whether or not terminology is a specific lexical feature of documentaries, will now be investigated. Werlich does not mention any specific type of vocabulary and neither does Gommlich, as illustrated in Section 1.2.1. Therefore, another text type classification, the influential analysis of Reiss and Vermeer (1984), has been related to Nichols' representation modes and the subject-based classification discussed in Section 1.1.

Although Nichols designed his representation modes for classifying non-fiction films from a filmic point of view, they can also be applied for analysing the verbal techniques used in the verbal modes of non-fiction. The table below provides a schematic overview of the verbal techniques used in the different modes. It is based upon Nichols' description of each mode and the examples he provides.

Table 3 The verbal techniques used in each representation mode

Verbal techniques -> Representation modes	Voice-over commentary		Interviews with talking heads	Spontaneous interviews	Staged interviews
	Voice-of-god	Voice-of--authority			
Expository	x	x	(x)	(x)	-
Observational	-	-	-	x	-
Interactive	-	x	x	x	-
Reflexive	-	x	x	-	x
Performative	x	-	-	x	x
Poetic	-	-	-	-	-

The above table illustrates that voice-over commentary (on or off-screen, as explained in 1.1.4.2) is the dominant verbal technique for explaining and informing, used prominently in the expository mode but also for the interactive, the reflexive and the performative

mode. Moreover, explaining and informing is the key element for determining terminology as a lexical feature in documentaries. As Franco et al. (2010, p. 24) put it, the voice-over narrator guides the audience's interpretation of an argument claiming reality through several responsible functions, like "linking scenes, introducing participants, explaining contexts, commenting on events, assessing people and situation". Hence, this authorial voice is associated by the viewer with expertise, reality and truth, the main characteristics of documentaries (see Section 1.1).

In the corpus of the present dissertation, made available by the Flemish public broadcaster, the commentary voice is represented by on-screen speakers ('voice-of-authority') and off-screen voices ('voice-of-god'). Traditionally in Flanders, on-screen speakers are subtitled while off-screen voices are translated with off-screen dubbing, the translation mode addressed in the present research.

To link voice-over commentary to terminology, in a first step, Reiss and Vermeer (1984) and in a second step Matamala (2010) and Cabré (2003) have been consulted. In their ground-breaking work *Towards a general theory of translational action* (1984), Reiss and Vermeer distinguish text type and text genre. Text type is a functional classification based on three basic communication forms: the informative text type (texts composed to relay news, knowledge, views), the expressive text type (when the information is consciously verbalized based on aesthetic criteria) and the operative text type (when the information offer conveys persuasively organised content in order to encourage the recipient to act in accordance with the intentions of the text sender). It must be said that the text types do not always occur in a pure form. An operative type can have an expressive secondary function or an informative type can have a secondary operative function and the three functions may even alternate in the same text. To the three basic communication forms, Reiss and Vermeer (1984) add a fourth one: the multimedial text type, a 'hyper-type' superimposed over the three others and consisting of written texts or oral speech with images or music. Documentary films contain images and usually also text, oral or written or both. They may also include diegetic and non-diegetic sound and music. For these reasons, documentaries are to be considered as a multimedial text type.

Nevertheless, the three basic text types can be distinguished as well in Nichols' representation modes. The expository and the interactive mode, offering information and representing an argument or an experience, each of them in a different rhetorical frame (see 1.1.3.1.), are informative text types, composed to relay news, knowledge and views. They may have a secondary operative function if the filmmaker seeks the audience's agreement. The observational mode (inviting the viewer to take an active position in understanding what is shown), the reflexive mode (inviting the viewer to be aware of the film process) and the performative mode (inviting the viewer to analyse and judge the subjective and affective dimensions of our knowledge) are operative types with a secondary informative function. Yet, all the modes can have an expressive secondary or

third function according to the way they are filmed and edited. The poetic mode[47], not considered here because of its few rhetorical elements, would mainly be an expressive type. However, being documentaries, i.e. recreating reality and claiming knowledge with images attached to the referent for which the viewer takes an assertive stance (see 1.1.2.), they will all have an informative component.

In summary, the verbal narration in documentaries known as 'voice-over commentary' is the dominant technique for informing. This informative function, which is one of the four text type classifications distinguished by Reiss and Vermeer (1984), is recognizable in all documentaries, as has been shown through Nichols' representation modes above and the definitions of documentaries in Section 1.1.2.

In the second step, the lexical features of informative text types will be analysed. Reiss and Vermeer explain that the language and style used in a text are particularly important for the classification of the text type. A high frequency of evaluative words (either positive or negative) and rhetorical elements like anaphora, hyperbole or leading questions may indicate an operative text type. Phenomena related to the 'principle of linkage' (e.g. rhymes, metaphors) may point to an expressive text type. If these features occur sporadically, or are missing altogether, the text type belongs to the informative type (1984). Reiss and Vermeer do not explain which linguistic features typify informative text types.

In order to examine whether or not the presence of terminology may constitute a lexical element of the informative type, Matamala (2010) and Cabré (2003) offer an interesting approach. Matamala, investigating terminology problems in the translation of documentaries through a case study with science documentaries, points out that science documentaries bring specialised topics through somewhat specialised speakers in a carefully chosen context, to a lay or a specialised audience. These four elements (topics, speakers, context and audience) offer a wide array of possibilities for organising an audiovisual product (2010). However, a documentary film will always popularise the specialised topic and aim at informing and entertaining the audience. Nevertheless, "science popularisation can be considered an instance of specialised discourse" (Matamala, 2010, p. 257). The four documentaries which the author analyses in her case study would be classified in the subject-based classification explained in Section 1.1 as 'science' (two of them) and 'nature' (the other two), whereas she calls them all 'science documentaries' considering nature phenomena and animals under the greater denominator 'science'. This classification is quite valid, the only reason for separating

[47] 'The poetic mode stresses mood, tone and affect much more than displays knowledge or acts of persuasion. The rhetorical element remains underdeveloped.' (Nichols 2001: 103)

'science' and 'nature' in the classification of the present thesis is the great number of nature documentaries in the VRT corpus.

In addition to this study about the presence of terminology in science documentaries, Cabré states in her article 'Theories of terminology', where she frames her Communicative Theory of Terminology:

> Terminological units have to be studied in the framework of specialised communication, which is characterised by such external conditions as sender, recipient and medium of communication, by conditions of *information treatment*, such as a precise categorisation determined externally by the conceptual structure, fixation and validated by the expert community, by specific and contextualised treatment of the topic, and finally, by conditions which restrict the function and objective of this communication. (Cabré, 2003, p. 188, my italics)

Under these conditions, Cabré (2003) explains that specialised discourse presents information through lexical and textual linguistic features. The lexical features are units belonging exclusively to the topic or having a limited meaning in this context despite their wider occurrence. The textual features organise the text in a concise and systematic way according to the structure of knowledge. These lexical features occupying "a node in the conceptual structure of a subject field and being semantically the minimal autonomous units of this structure" constitute the terminological units (Cabré, 2003, p. 189).

Considering terminological units as sets of conditions which distinguish them from other units, Cabré indicates three areas that determine the conditions: from the perspective of the cognitive component of the units, from the communicative component and from the linguistic component.

From the perspective of the cognitive component, term units depend on a thematic context, occupy a precise place in a conceptual structure and their meaning is explicitly fixed and determined by their place in the structure.

From the perspective of the communicative component, term units occur in specialised discourse to which they adapt according to their thematic and functional characteristics. As they are acquired through a learning process, they are handled by specialists in the field.

From a linguistic point of view, which is of interest in this line of thought, terminological units have, inter alia, a lexical and syntactic structure. They occur as nouns, verbs, adjectives or adverbs and their meaning is discreet with a special subject (2003).

This theory explains how terminology exists in our language system and provides the information required for the present research that term units are lexical features in specialised discourse, which links voice-over commentary to terminology.

In summary, Cabré's analysis of terminological units and Matamala's study of science documentaries have shown that under certain conditions of information treatment and communication, the lexical features in specialised discourse - including what is referred to as 'science and nature' documentaries in the corpus under study - constitute the term units. As a consequence, it is fair to say that, under certain conditions, terminology is a linguistic feature of informative text types.
In Section 1.3, the concept and definitions of terminology will be addressed in greater detail.

1.2.3 Audiovisual translation and terminology

The translation of terminology is a specific challenge in all text types. First of all, the translator has to identify the term in the source text. Next, the meaning and the context have to be understood. In the case of the translation of documentaries for television or the cinema, the translator has to take into account the heterogeneous target audience in terms of age, cultural background and expertise. Therefore, terminology translation occupies a prominent position in the translational choices of audiovisual documentary translators (Franco, 2000; Espasa, 2004).

Matamala explains that terminological problems include "identifying terms, understanding terms, finding the right equivalent, dealing with the absence of, or the inability to find, an adequate equivalent, dealing with denominative variation, choosing between 'in vivo' and 'in vitro' terminology and avoiding wrong transcriptions" (2010, p. 259). 'In vivo' and 'in vitro' are techniques used in biology, meaning 'in the living body of a plant or animal' (in vivo) vs. 'in a test tube' (in vitro)[48]. This distinction, which was proposed by Cabré (1999) and further discussed in Matamala (2009) refers to the fact that the translator sometimes has to choose between an equivalent proposed by linguistic authorities and found in terminological or lexicographical works (in vitro) or a term which is really used by specialists in the field, which might be a loanword (in vivo) e.g. *Brout-Englert-Higgsdeeltje (in vitro) vs. higgsboson (in vivo) (The Hunt for the Higgs, 2012)*.

It is for all these reasons that the hypothesis of the present thesis, as described in the introduction, is that it must be beneficial to translation efficacy to support audiovisual translators in their work with a domain-specific, bilingual glossary. The glossary is English-Dutch (see Chapter 2), one for each documentary used in the experiments, drawn up by a terminology extraction system. Chapter 4 explains the experimental set-up and the results, Section 1.3 deals with terminology extraction systems.

[48] https://www.merriam-webster.com/dictionary

To date, such systems are mainly used for texts with a high degree of repetition and a large amount of terminology such as technical, scientific, financial and legal texts (Christensen & Schjoldager, 2010; Lagoudaki, 2010). One of the issues this study will look into is the extent to which the method is useful in the context of AVT of documentaries.

Matamala (2010) argues that science documentaries can be considered as specialised discourse (see also 1.2.3) and a preliminary viewing of the films in the corpus under study (see Chapter 2) confirms that some documentaries contain a mixture of general utterances and domain-specific terminology while others contain more and recurring terms. One question is how much terminology they contain and another, whether or not this terminology is specific enough to be detected by automatic terminology extraction systems. These issues will be further explored in Chapter 3 'The methodology research'.

1.2.4 Audiovisual translation and technology

The automation of the translation of audiovisual content is an ever-growing research field. Several projects mentioned below have been carried out in the last decade, all of them with the intention to increase productivity, reduce costs and enhance the quality of translation results through the introduction of technologies, in an attempt to meet the market demands. Unfortunately, no publications detailing the results of the projects mentioned below or their shortcomings appear to be available[49].

The EU-BRIDGE[50] project was concluded in January 2015. EU-BRIDGE advanced automatic speech-recognition and speech-translation technology (Large Vocabulary Continuous Speech Recognition combined in sequence with Statistical Machine Translation, technical details are presented in the technology catalogue[51]) that can convert speech from lectures, meetings and telephone conversations into other European and non-European languages. Academics, together with engineering and business experts have focused on usage in four use cases: automated captioning of broadcast news, simultaneous translation of university lectures, speech translation services for the European Parliament and simultaneous translation for webinars. The prospective users are European companies operating on the audiovisual market (in particular TV captioning and translation). The final report states that core technologies of speech translation systems have been enhanced and offered to application developers and that the technology has been inserted into the market of the four use cases.

[49] Except for the SUMAT-project: the researchers kindly sent me the extended final report.

[50] http://www.eu-bridge.eu/ (accessed 12/03/2015) *European Union grant agreement n°287658.*

[51] http://www.eu-bridge.eu/img/fs_eubridgetech_screen.pdf (accessed 23/06/2015).

The SUMAT-project[52] (Subtitling by Machine Translation) wrote its final report in 2014. SUMAT introduced statistical MT techniques into the subtitle translation processes in order to develop an online subtitle translation service for 9 different European languages combined into 14 different language pairs. This has resulted in large-scale semi-automatised translation of subtitles of both freelance translators and subtitling companies. Overall, the SUMAT project has developed a state-of-the-art machine translation technology for subtitling, along with precise evaluations of its potential and current limitations. According to the final report, the technology is viable as it stands, given the quality ratings and the productivity gains achieved. However, assessing the quality and the usefulness of MT for subtitling through evaluation of the post-editing task by professional subtitlers, it was revealed that three core aspects need to be improved:

1) The quality of MT (although judged sometimes surprisingly good)
2) The automatic quality estimation and filtering of MT output (reducing the cognitive effort needed for human estimation of the machine translated output and allowing subtitlers to focus on the less frustrating parts of the post-editing task)
3) The adaptation of the post-editing user-interfaces to enable efficient correction of the most typical MT errors.

The inEvent project - Accessing Dynamic Networked Multimedia Events - was organised[53] between 2011 and 2014. The main goal of inEvent was to develop new means to structure, retrieve, and share large archives of networked, and dynamically changing, multimedia recordings, mainly consisting of meetings, video-conferences, and lectures. According to the final public report[54], innovative progress was made. Visual recognition technology was developed to segment and classify the narrative structure of video. Deep neural networks were employed for automatic audio transcription. Natural language processing was used to qualify and quantify comments made on TEDs videos and to recommend other potentially interesting videos to users. Advances in state-of-the-art technology were used to identify who said what and to link the speakers across different recordings.

During the same period (2011-2014), the TOSCA-MP (Task-oriented search and content annotation for media production) project was set up. It aimed to develop user-centric content annotation and search tools for professionals in networked media production and archiving (television, radio, online), address their specific use cases and workflow requirements. The TOSCA-MP results enable professionals in media production and

[52] http://cordis.europa.eu/project/rcn/191741_en.html (accessed 23/06/2015).

[53] http://www.inevent-project.eu/

[54] http://www.inevent-project.eu/files/inevent-final-public-report (accessed 23/06/2015).

archiving to seamlessly access content and indexes from distributed heterogeneous repositories in the network. This is achieved by providing technologies that allow instant access to a large network of distributed multimedia databases, including state-of-the-art metadata linking and alignment. The distributed repositories can be accessed through a single user interface that provides novel methods for result presentation, semi-automatic annotation and means of providing implicit user feedback[55].

TransLectures[56] was an EU-funded project that ran from 2011 to 2014, to develop innovative, cost-effective tools for the automatic transcription and translation of online educational videos. The project was based on automatic speech recognition (ASR) and machine translation (MT) technology and was evaluated in a real-life context. According to the publishable summary report[57], quality improvements have been achieved for the transcription and translation of the languages involved, the massive adaptation of acoustic tools and language models has improved ASR systems and MT techniques, intelligent interaction techniques for transcription and translation have been integrated into the different ASR and MT toolkits and the tools for integrating the transLectures Platform into the Opencast Matterhorn platform have been developed, tested and publicly released. The transLectures solutions can be adopted by educational repositories to overcome language barriers.

uDialogue[58] is an ongoing Japan-based project, supported by Core Research for Evolutionary Science and Technology (CREST). The idea is to create an environment where individuals can freely and naturally transmit and receive information via speech signals, bridging the gap between spoken communication and advanced network telecommunication devices, unevenly distributed across the globe and creating a flexible speech technology that considers the domains of facial expression, gesture, speech quality, timing, and similar elements of communication. The objective is to establish a system in which a user can easily create speech-based interactive content including numerous spoken dialogues that they can evaluate.

To conclude, the SAVAS-project[59] seeks to develop an Automatic Speech Recognition (ASR) technology for multilingual live subtitling, specifically tuned to the needs of the broadcasting and new media industries.

What all these projects have in common is that they turn to state-of-the-art technologies for improving the efficiency of specific forms of audiovisual translation or the searchability of media content and databases. The projects dealing with AVT mainly

[55] http://tosca-mp.eu/ (accessed 23/06/2015).

[56] https://translectures.eu//

[57]https://www.translectures.eu/wp-content/uploads/2015/01/transLectures-T36-15Dec2014-publishable-summary.pdf

[58] http://www.udialogue.org/

[59] http://www.fp7-savas.eu/ (accessed 23 September 2015).

consider subtitling as translation mode and speech recognition and machine translation as technological support systems. They provide evidence for the view that productivity can be increased and costs reduced while enhancing the quality of the translation and introducing technology for the translation of audiovisual content. The present thesis supports this view, as it considers research from a different angle: the application of terminology extraction systems to a translation context of off-screen dubbing for documentaries. This application will be investigated in order to verify the impact of bilingual glossaries on the translator's workload and workflow.

1.3 Insights from terminology extraction studies

Since the third element of this dissertation's main research question, besides documentaries and their translation, discussed in the previous sections, is the bilingual glossary and, consequently, terminology, it is useful to address the main aspects of terminology and terminology extraction.

1.3.1 Definitions of terminology

Definitions of terminology all encompass the same aspects: concepts, terms and specialised domain/language. According to Bowker, "terminology is concerned with the naming of concepts in specialised domains of knowledge" (2008, p. 286). Pavel and Nolet define terminology as "the scientific study of the concepts and terms used in specialised language" (2001, p. xvii).

An interesting and exhaustive definition is offered by Sager who explains that the word 'terminology' has three meanings:

1) The set of practices and methods used for the collection, description and presentation of terms[60]
2) A theory, i.e. the set of premises, arguments and conclusions required for explaining the relationships between concepts and terms which are fundamental for coherent activity mentioned under above point 1)
3) A vocabulary of a special subject field (1990, p. 3)

[60] This first aspect is also known as 'terminography' or 'terminology work' or 'terminology management' (Bowker, 2008; Allard, 2012) and regards applied linguistics, computational linguistics and computer science (Zielinski & Ramirez, 2005).

Issues related to the second and the third aspect of this definition i.e. identifying the nature of terms and terminology will be addressed in the section below. Next, the first aspect will be tackled providing an overview of current research and surveys regarding (the application of) terminology. Finally, the different approaches that have been proposed for automatic terminology extraction will be discussed. The literature overview is concluded by a description of the term extraction systems which were used for this thesis..

1.3.1.1 Concepts and terms

Sager's second meaning of terminology deals with concepts and terms. Concepts are units of thought used to organise our knowledge. They are related to other concepts, making up the knowledge structure of a domain. The definition of a concept provides the bridge between the concept and the term used to designate it. A terminological definition differentiates a concept and its associated term from other concept-term units (Bowker, 2008). In other words, terms are the linguistic designations "assigned to concepts used in special languages that occur in subject-field or domain-related texts" (Wright, 1997, p. 13) or as Bowker (2008, p. 286) puts it: "terms refer to discrete conceptual entities, properties, activities or relations that constitute knowledge in a particular domain".

They occur as single-word units or as multi-word units, depending on language-specific conventions. Germanic languages for instance, combine word elements, whereas Romance languages use linking elements to form multi-word units. Each component is itself a single-word term that can be defined on its own, yet the multi-word term constitutes one concept representing a greater whole (Kageura, 2015).

The relationship between term and terminology is straightforward. As far as the third aspect of Sager's above-cited definition is concerned, one can conclude that collectively, terms form the terminology of a discipline.

1.3.1.2 Terminology management

The first aspect of terminology Sager (1990) mentions in his definition, "the set of practices and methods used for the collection, description and presentation of terms", covers areas of applied linguistics, including specialised translation. In fact, translators of specialised texts need to identify the terms in the source text and find the correct equivalent in the target language. This search can be time-consuming and labour-intensive. Moreover, as described in the previous Section 1.2, the translation of terminology is a specific challenge in all text types as it also occupies a prominent position in the translational choices of audiovisual documentary translators.

The way terms are collected, described and presented differs according to the scope and the technology at hand. Bowker (2011, 2015) argues convincingly that there is a growing divide between the terminology work of terminologists and that of translators.

Terminologists gave up print-based in favour of electronic corpora and have now corpus-based tools at their disposal to interrogate the huge amount of texts. Their target users are in-house translation teams, their term records are prescriptive, with a view to standardization and cover a broad range of domains. The recorded units are terms in the canonical form and the term bank is usually not integrated with other tools, e.g. the well-known EU's inter-institutional, multilingual terminology database IATE[61] and the GDT[62], the domain specific terminology database of the 'Office québécois' in French and English.

Translators, however, can integrate their term base into Translation Environment Tools (TEnTs). This is a set of software programs that provide an integrated framework for the support of translators, basically a translation memory and a terminology database or term base directly interacting with word processors and term extractors, sometimes with machine translation systems (Bowker 2008 and 2015, Sager 1990). The target user of the term base is the individual translator. The term records are descriptive, product-oriented and they are limited to one domain, so per client and per domain a separate data base. As translators are concerned with all lexical items that pose a challenge, even if they are not terms in the strict sense but occur frequently e.g. expressions or words that can lead to (spelling) errors, terms are not necessarily recorded in the canonical way (including not only the foreign-language equivalent but also grammatical information, synonyms, a definition, context, usage observations, sources etc.). Translators may even simply prefer to generate searches 'on the fly' as the need arises without producing any term record but just a list of equivalents (Bowker, 2011; 2015).

Whether or not terminology management is done by terminologists or by translators, keeping track of terminological information nowadays is important, given that the number of documents in electric form is overwhelming and the time-to-market pressure intense.

In order to identify user practices regarding terminology management, Allard (2012) surveyed 104 users of term bases that were integrated with TEnTs. Of the respondents, 74% were translators, while the others were terminologists, revisers and project managers. She demonstrated that almost half of the respondents considered that terminology management is important in order to:

- record expressions and their equivalents that require extensive research.
- create subject-specific glossaries that help make translational choices.

In a case study for a large biotechnology company involved in the translation of patent applications, Coombs (2014) showed that terminology management is one of the best practices to improve consistency, reduce error risk and reduce time to grant patents and

[61] http://iate.europa.eu/switchLang.do?success=mainPage&lang=en

[62] http://www.granddictionnaire.com/index.aspx

increase the number of patent filings. Establishing a common terminology database for all translation service providers who collaborate with the enterprise improved the translation quality significantly and maximized the return of investment.

Similar results concerning quality and speed were found by Warburton (2013) and Kelly and De Palma (2009). Warburton proposed a method to integrate terminology extraction in the translation pipeline, something which is particularly beneficial for companies that translate large, terminologically-rich corpora. Kelly and De Palma interviewed 'veterans' in translation, as well as newcomers working for companies in Europe and North America, to determine the value of terminology management and found that investing time and money in terminology management improves quality, is time efficient, minimizes corrections and strengthens a company's brand. The return of investment of terminology management has been interrogated as well, although it is hard to separate the terminology work from the translation process as it forms an integral part of the translation-revision cycle.

Champagne (2004) analysed the working process of twelve companies in order to establish the economic value of terminology. He conducted interviews to provide a snapshot of Canada's terminology market, mailed questionnaires to gather data concerning the terminology work in the company and organised focus groups to validate responses. In terms of time, he found that terminology occupied 15 to 30% of the working time of translators and revisers. This is an average of 11h20 per week, which is in line with the survey of Allard, who indicates an average of 9h42 per week dedicated to terminology management (2012, p. 124). However, LISA's survey about trends in terminology management within the localisation industry indicated that half of the respondents spend no more than 3h48 per week on terminology-related work (Lommel, 2005, p. 2). The difference in these results could be due to the fact that clients and managers in the localisation industry still do not perceive terminology as very important, so fewer resources may be invested in specific tools and consequently, less time in terminology work (Allard, 2012; Lommel, 2005). In terms of productivity, Champagne concluded that terminology raises productivity in writing, translation and revision by 20% and increases the profit margin by 10%. These data suggest that even though terminology plays a vital role in the language process, it is not seen in terms of economic value but as a way to make the work more productive and of better quality.

In a more recent survey, Herwartz (2011) calculated the costs and benefits a company incurs when working with and without terminology management. She also conducted interviews to gauge the efforts for review purposes, the time for research, coordination and approval of terminology and the rate of reuse of texts. Her findings showed that savings vary from a mere 5 to 50%. Therefore, the study recommends newcomers in terminology management to search for benchmarking with experienced companies to cut costs. Herwartz also found that monitoring terminology work demonstrates to the company how effectively the terminology is managed.

The above-mentioned studies all prove that, besides return on investment, the overall benefit of managing terminology lies in more terminological consistency in the final product and, as a consequence, improved quality (Champagne, 2004; Coombs, 2014; Kelly & De Palma, 2009; Lommel, 2005). However, they investigate terminology management by means of surveys and case studies i.e. measurements after the translation process, rather than as part of the translation process.

The present thesis is a first step towards filling this gap. Translation experiments will be conducted in order to monitor and measure the terminology research during the translation process itself. In order to determine the added value of terminological support in the translation process, Master's students in translation and professionals were asked to translate extractions of off-screen dubbing for documentaries in three experimental settings, viz. without a term base, with a manually extracted term base and with an automatically extracted term base.

1.3.2 Terminology extraction

Frantzi, Ananiadou and Tsujii (1997) offer the following short but useful definition of terminology extraction (or term recognition or term mining, as this practice is also called in the literature): "Automatic term recognition is the extraction of technical terms from special language corpora with the use of computers" (1997, p. 1). Back in 1996, Kageura and Umino stated that "[a]utomatic term recognition in particular is much needed because a simple but coherently built terminology is the starting point of many applications such as human or machine translation [...] and because manual efforts cannot keep up with the rapid growth of technical terms" (p. 259).

Different strategies to detect terms, ranging from linguistic patterns to statistical filters or a combination of both, have been developed, tested and evaluated in the past decades (L'Homme, Heid, & Sager, 2004). However, manual verification of the output is still needed. This section provides an overview of the main approaches and their characteristics.

1.3.2.1 Monolingual terminology extraction

- THE LINGUISTIC APPROACH

The linguistic approach starts from the morphological and syntactic structure of terms i.e. characteristics of term formation patterns, which are expressed as part-of-speech code sequences (e.g. N N, N prep N, Adj N). This means that linguistically-based systems are always language dependent. In English and Dutch for example (the languages under consideration for the experiments in this study) multi-word terms are mainly compounded by combining nouns. However, differences can be observed. While in Dutch,

the compounds are often not separated by white spaces (forming a single-word term), in English they are (constituting a multi-word term). In Romance languages like French and Italian, prepositional phrases and noun-preposition-noun are the most frequent compounding strategies (Macken, 2010; Macken, Lefever, & Hoste, 2013). The linguistic approach relies on a corpus that is first automatically processed by a so-called Part-of-Speech tagger which assigns morphosyntactic information to all words in the given corpus. The third column in Table 4 exemplifies such a PoS output for one of the sentences in our corpus.

In addition, the corpus is also lemmatized (as shown in the second column in Table 4), which means that for each word, the base form, or lemma, is generated. For conjugated verbs, the infinitive is produced, for plural nouns the singular form, for adjectives the stem etc.

After lemmatization and PoS tagging, syntactically related consecutive words are grouped into chunks on the basis of a superficial sentence analysis (Column 4 in Table 4). The chunker used for the present experiment tries to form constituents on the basis of a set of constituency and distituency rules. Constituency rules define which parts-of-speech can consequently occur within a constituent, while distituency rules define which parts-of-speech cannot occur adjacent to each other within a constituent (Macken, 2010). Other chunkers might work differently with other sets of rules. For instance, the recursive chunker for Dutch proposed by Spranger (2003) includes chunks or phrases of a given category within another chunk of the same category e.g. *[NP [AP [PP door de rechter PP] benoemde AP] curatoren NP].*

Table 4 Example of PoS tagging and chunking

Word	Lemma	PoS	Chunk
Planet	Planet	NNP	B-NP
Earth	Earth	NNP	I-NP
is	be	VBZ	B-VP
unique	unique	JJ	B-AP
.	.	.	O
The	the	DT	B-NP
latest	late	JJS	I-NP
information	information	NN	I-NP
from	from	IN	B-PP
satellite	satellite	NN	B-NP
maps	map	NNS	I-NP
,	,	,	O
sonar	sonar	NN	B-NP
and	and	CC	O
radar	radar	NN	B-NP
images	image	NNS	I-NP
have	have	VBP	B-VP
been	be	VBN	I-VP
brought	bring	VBN	I-VP
together	together	RB	B-ADVP
to	to	TO	B-VP
create	create	VB	I-VP
this	this	DT	O
.	.	.	O

Key for PoS codes:

CC coordinating conjunction
DT determiner
IN preposition or subordinate conjunction
JJ adjective or ordinal numeral
JJS adjective or ordinal numeral, superlative
NN noun
NNP proper noun singular
NNS plural common noun
RB adverb
TO preposition/infinitive marker
VB infinitive
VBN verb, past participle
VBP verb, present tense (not 3p)
VBZ verb, present tense (3p)

Key for chunking codes:

A B I O chunking classifier (one of the more common types of chunking classifiers), labels each token as being the Beginning of (B), the Inside of (I) or entirely Outside (O) of a span of interest.

NP noun phrase
VP verb phrase
AP adjective phrase
ADVP adverbial phrase

As terms often do not present linguistic characteristics that distinguish them from non-terms, a purely linguistic terminology extraction system tends to overgenerate.

- THE STATISTICAL APPROACH

A second approach to automatic terminology extraction is the statistical approach. It is language independent and is based on quantifiable characteristics of term usage. One such characteristic is that terms tend to occur more frequently in specialised texts than in general domain texts. The system filters term candidates by applying different statistical measures. A threshold is then established to filter out valid terms (Zielinski & Ramirez, 2005; Macken et al., 2013).

In a statistical approach, the candidate terms are called n-grams: an n-gram of length 1 consists of one word, an n-gram of length 2 consists of two words etc. High frequency words and function words (e.g. the, and, of, from, this, when etc.) considered as not

belonging to specialised language are excluded from the list of candidate terms and often included in a stop-word list.

The candidate terms are then filtered by applying two statistical metrics: termhood and unithood. Kageura and Umino explain that unithood "refers to the degree of strength or stability of syntagmatic combinations or collocations" and termhood "refers to the degree that a linguistic unit is related to [...] domain-specific concepts" (1996, p. 260-261).

Unithood is relevant to complex terms and other complex units as grammatical collocations and idiomatic expressions as long as the multi-word units present a high degree of cohesiveness. Moreover, multi-word units are considered to be highly prevalent in technical domains (Bourigault & Jacquemin, 1999; Heylen & De Hertog, 2015; Kaguera & Umino, 1996). An example of a multi-word unit with a high degree of cohesiveness is 'free-floating blocks of concrete' (*The earth machine,* 2011). It is considered to be one term and so is its corresponding term in Dutch 'vrij bewegende betonnen structuur'. In addition to that, it is important to keep in mind that single-word units are less specified than their multi-word counterparts. In the work of Nakagawa & Mori (2002) it was found that 85% of terms are compound nouns. Bourigault and Jacquemin (1999), for example, stated that single word terms are too polysemous and too generic while multi word terms represent fine concepts and are semantically more specified e.g. *facial nerve paralysis* is more specified that *paralysis.*

Termhood means that each entry in the extracted lexicon, whether it is a simple or a complex linguistic unit, should refer to an object or action that is relevant to the domain. This means that the relative frequency of the extracted linguistic unit in the given domain will be higher than in other domains or general speech (Kageura & Umino, 1996). For instance, the word 'bird' in a documentary about the flora and fauna in Madagascar might be considered to be a term, while the same word in a documentary about the creation of the earth, is not.

Several statistical filters, measuring termhood and unithood, have been developed to distinguish terms from non-terms among the candidate terms: the C-Value (Frantzi et al., 2000), the Log-likelihood ratio (Dunning, 1993) and the TF IDF[63] (Spela Vintar, 2004) amongst others. Frantzi et al. (2000) describe C-Value as a statistical measure to assign termhood to a candidate string and focus on the extraction of nested collocations by analysing the frequencies of a collocation used as a part of a longer collocation (Frantzi et al., 2000). The Log-Likelihood Ratio is a statistical test that can determine whether or not the differences in relative frequency (see paragraph below) of a candidate terms in two different corpora is statistically significant (Rayson & Garside, 2000). The TF IDF filter (term frequency / inverse document frequency) is a way to score the importance of terms

[63] Not applied in TExSIS, the system used for the present research, but mentioned because of its well-known usage in information retrieval, but also in terminolgy extraction.

in a document based on how frequently they appear across multiple documents (Spela Vintar, 2004). Chapter 4 will elaborate on these filters.

The underlying principle of these filters is that terms belong to a specific domain and occur more frequently in that domain than in common language. Formally, frequency is expressed in absolute frequency (the number of times a word occurs) and in relative frequency (the number of times a word occurs compared to the total number of words in a corpus).

A purely statistical terminology extraction system tends to undergenerate, because frequency filters often fail to detect new terms (that occur infrequently). However, they also tend to overgenerate as frequently occurring linguistic units will be extracted as well.

- THE HYBRID APPROACH

Hybrid systems combine both linguistic and statistical approaches. Candidate terms are first extracted on the basis of linguistic methods and then filtered by applying statistical scores. As linguistically-based systems tend to overgenerate and statistical methods tend to undergenerate - producing however some noise as well, - most state-of-the-art systems use hybrid approaches (Macken et al., 2013).

The first researcher to develop a hybrid system was Béatrice Daille. She evaluated several statistical scores, such as frequency, association scores, diversity and distance metrics, discovering that frequency undoubtedly characterizes terms and that incorporating linguistics in a statistical system increases the accuracy of the extraction of lexical resources (1996).

1.3.2.2 Bilingual terminology extraction

While terminology extraction can be carried out in a monolingual setting, also bilingual systems have been developed. These systems are particularly interesting for translational purposes. A parallel corpus consisting of source texts and their translation is the starting point for finding translational equivalents.

The architecture of a typical bilingual terminology extraction system is as follows. First, monolingual term extraction on both languages is carried out separately, which combines the linguistic and the statistical approach as explained in part 1.3.4.1. Sometimes, both corpora are pre-processed by means of a lemmatizer, PoS tagger and chunker. Next, a list of candidate terms is extracted and then, terms are filtered out of these candidate terms. Doing so, a monolingual term list is obtained for each language. Separately, sentence alignment software first aligns the sentences of the bilingual corpus and then a word alignment system links – where possible – the corresponding words (Macken et al., 2013).

- SENTENCE ALIGNMENT

Sentence alignment is the process of identifying equivalent sentences or sentence chunks in parallel texts. A range of models have been developed for this task, mainly based on two different approaches: the sentence-length-based approach and the lexicon-based approach.

The sentence-length-based approach was introduced by Gale and Church (1991). Its guiding assumption is that longer sentences - expressed in number of characters - in the source language tend to be translated into longer sentences in the target language, and shorter sentences into shorter, which makes this approach language-independent. A probabilistic score is assigned to each proposed correspondence of sentences in order to find the maximum likelihood alignment of sentences (Gale & Church, 1991).

The lexicon-based approach was elaborated by Kay and Röscheisen (1993). The guiding assumption for their algorithm is that if sentences are each other's translation, the corresponding words in the sentences must be translations as well. In other words, a pair of sentences containing an aligned pair of words must themselves be aligned. The Kay and Röscheisen algorithm performs two functions simultaneaously i.e. sentence alignment as recorded in the sentence alignment table[64] and word alignment as recorded in the word alignment table[65], which is in effect a probabilistic dictionary. The algorithm only depends on information derived from the texts themselves, no other information about the language pair is involved (McEnery et al., 2006).

- WORD ALIGNMENT

Word alignment is the process of statistically matching up words within pairs of aligned sentences. It is based on "the assumption of co-occurrence: words that are translations of each other co-occur more often in aligned sentence pairs than that they occur randomly" (Macken, 2010, p. 21). A good word alignment is important for bilingual terminology extraction (Brown, Della Pietra, Della Pietra & Mercer, 1993).

The most common approach to word alignment is the generative IBM translation models, described by Brown et al. (1993), which assign a probability to each of the possible word-by-word alignments for a pair of sentences that are translations of each other (Brown et al., 1993). The guiding principle is the expectation-maximization algorithm used in statistics. This is an iterative learning method that fills in the gaps of incomplete

[64] A table that records for each pair of sentences how many times the sentences were set in correspondence with the algorithm (Kay & Röscheisen, 1993).

[65] A list of pairs of words, together with similarities and frequencies in their respective texts, that have been aligned by compairing their distribution in the texts (Kay & Röscheisen, 1993).

data and trains a model in alternating steps (Koehn, 2009). In the expectation step, the model is applied to the data, filling in the gaps with the most likely alignments. In the maximization step, the model learns from the data which are augmented with all possible guesses for the gaps and weighted with their corresponding probabilities. The IBM models take into account word frequencies, word order and the probability that one source word aligns to more than one target word. The models iterate from step one and two until convergence (Koehn, 2009).

Given the minimal linguistic content they have, learning as they do from the data itself, it is reasonable to argue that word-by-word alignments will perform better as more data is available. However, the corpus under study (described in Chapter 2 'Corpus') is rather limited. Consequently, some terms do not feature in the automatically extracted glossary, because the system had insufficient data to detect them e.g. terms occurring only once or twice, like flycatcher (the system only detected the term in English but could not detect the Dutch translation 'paradijsmonarch') or multi-word terms translated by a single-word term like, giraffe-necked weevil (the system only matched the first word 'giraffe' with 'giraffenkever') (*Madagascar -The Island of Marvels*, 2011).

1.3.2.3 Bilingual terminology extraction systems

Most common approaches to terminology extraction first identify term candidates monolingually before aligning source and target terms. Other systems take a bilingual perspective from the start (Tiedemann, 2001; Vintar, 2010).

For the purpose of this thesis, three bilingual systems were tested on our corpus of off-screen dubbing for documentaries (presented in Chapter 2) in order to select the best performing system to be used for the experiments: one of the major commercial systems SDL Multiterm Extract 2011 Trados®[66], a free-ware system Similis®[67] and TExSIS, a system developed by the University of Ghent[68] (Macken et al., 2013).

The systems have different underlying technologies. SDL Multiterm Extract 2011 Trados® does an automatic mono- and bilingual term identification and extraction, based on a statistical approach and bilingual concordance to show the occurrence of the term in context[69]. Similis®, a hybrid system, runs an initial linguistic analysis that looks at the sentences and identifies the chunks (usually nominal or verbal groups, but not always syntactic constituents). The second stage of analysis uses algorithms to identify the basic form of each word of the chunks and their grammatical categories (Planas, 2005). TExSIS

[66] http://www.sdl.com/cxc/language/terminology-management/multiterm/
[67] http://similis.org/linguaetmachina.www/index.php
[68] http://www.lt3.ugent.be/en/projects/texsis/
[69] http://www.translationzone.com/products/sdl-multiterm/extract/index-tab2.html#tabs

is a subsentential alignment system that starts from sentence aligned parallel texts. In a first step, it generates anchor chunks (viz. chunks that can be linked with a very high precision based on lexical correspondences and syntactic similarity). In a second step, a bootstrapping approach is used to extract language-pair specific translation rules. The motivation behind this two-step system that links linguistically motivated phrases, is that technical texts use a lot of technical terms consisting mainly of multi-word units (see Section 1.3.2.1 'Statistical approach'). Candidate terms are then generated from linguistically motivated aligned chunks and statistical filters are applied to determine the specificity of the candidate terms (Macken et al., 2013).

The evaluation of the output of the three systems applied to the corpus under study is discussed in Chapter 3 'Methodology'.

First, the corpus will be presented in Chapter 2.

Chapter 2
COMPOSITION OF THE RESEARCH CORPORA

2.1 Introduction

The Flemish public broadcaster VRT generously made available a corpus of scripts and their translation (Section 2.4 below provides more information about the VRT guidelines for translation) consisting of all documentaries broadcast between 2005 and 2013. The translations, intended for a Flemish target audience and culture, were mainly from English into Dutch, which is the major language pair not only for documentaries on Flemish television but also for foreign films and programmes on Flemish television in general. For this reason, English-Dutch is the language combination investigated in the present dissertation.

In a subtitle country like Belgium, the main translation modes for documentaries are subtitling and off-screen dubbing, each of which create their own specific challenges in terms of language register and technical skills on the part of the translator (see Section 1.2). In this study, only off-screen dubbing will be covered, as subtitling for documentaries has been the subject of many studies e.g. Franco 1999, Kaufmann 2004 and 2008, Remael 2003 and 2007, Taylor 2002. More details about these studies are offered in Section 1.2.1. Moreover, the use of translation technology for the automation of subtitles for many different kinds of audiovisual texts has been investigated in several projects (see Section 1.2.4.). In contrast, the use of translation technology for off-screen dubbing appears to have been addressed only recently by the ALST-project (Matamala, Fernàndez-Torné and Ortiz-Boix 2012), that investigates the application of machine translation and post-editing to off-screen dubbing for wildlife documentaries. The present study also addresses off-screen dubbing for documentaries, narrowing the focus on the use of terminology extraction systems for this audiovisual translation mode.

With respect to the type of documentaries examined in this research, the corpus is limited to natural-science documentaries, because these episodes were expected to

contain recurring, domain-specific terminology to be considered for accurate, automatic term detection, as explained in Sections 1.2 and 1.3.

The VRT corpus contains 181 episodes of natural-science documentaries in English. These films could be divided roughly into three subject-based categories: wildlife (52%), earth and space (34%) and human body (14%). In order to compose the specific research corpora – one for the preparatory study and one for the experiments – a selection was made in which all three categories were equally represented.

2.2 The preparatory corpus

Section 1.2.3. demonstrated that natural science documentaries can be considered as specialised discourse and that specialised discourse presents information through lexical features constituting terminological units. Therefore, a preparatory study was needed in order to investigate whether or not the corpus under study contains terminological units.

The preparatory study was organised as follows: ten films proportionally distributed by subject were selected, called 'the preparatory corpus'. Table 5 below presents the ten documentaries and their subjects. All source-text and target-text sentences were aligned by means of a translation memory software Similis®[1] and the alignment was verified manually. Next, one annotator labelled all terminological units manually, taking into account the criteria of termhood and unithood explained in Section 1.3.4.1. A comparison expressed in percentage was subsequently made between the number of term types (= unique terms, repetitions excluded) and the number of unique words. In Chapter 3 on 'Research methodology', the preparatory study is discussed.

[1] http://similis.org/linguaetmachina.www/index.php

Table 5 The preparatory corpus

Title	Channel	Category
Around the world in 80 gardens – Australia/New Zealand (2008)	Eén	Wildlife
Big cat week (2006)	Eén	Wildlife
Madagascar – Island of marvels (2011)	Canvas	Wildlife
Polar bear – Spy on the ice (2011)	Canvas	Wildlife
The natural world – Empire of the desert ants (2010)	Canvas	Wildlife
How earth made us – Deep earth (2010)	Canvas	Earth and space
Is the magnetic pole about to flip? (2010)	Canvas	Earth and space
Earth machine – Land (2011)	Canvas	Earth and space
E-numbers, an edible adventure (2010)	Canvas	Human body
Horizon - The secret world of pain (2011)	Canvas	Human body

2.3 The experiment corpus

A second corpus, called the experiment corpus (see Table 6), was constituted in order to investigate the core of the present dissertation i.e. to understand whether or not the integration of automatically extracted bilingual glossaries in the translation process reduces the workload of documentary translators. For this purpose, translation experiments have been conducted, involving Master's students in translation and professional translators.

The experiment corpus was created as follows: out of the preparatory corpus, three representative documentaries in terms of subject (one of each: wildlife, earth and space, human body) were selected. The selection was made on the basis of their domain-specific terminology, giving priority to the specificity and not to the frequency of the terms as will be explained in Chapter 3. This corpus was used firstly to test three automatic terminology extraction systems (Methods and results of the test are discussed in Chapter 3). Secondly, the experiment corpus served to draw up a manually-labelled and an automatically-extracted bilingual glossary for each episode which the translators used for the experiments (discussed in Chapter 4). Thirdly, excerpts from this corpus served as the source text for all the translation experiments.

Table 6 The experiment corpus

Madagascar – Island of marvels (2011)	Wildlife
Earth machine – Land (2011)	Earth and space
Horizon - The secret world of pain (2011)	Human body

2.4 VRT off-screen dubbing

As the entire corpus was made available by VRT, all the translations in the corpus were done by VRT translators following VRT guidelines. It is, thus, useful to have a closer look at these guidelines.

VRT has drawn up specific guidelines for off-screen dubbing. The guidelines are not particularly 'objective' in the way they approach their source texts. Instead, they pander to the VRT style and deal with the perceived specificities of the source texts, i.e. the verbal channel of the documentaries, from that perspective. The guidelines for the translation of voice-over commentary viz. off-screen dubbing (see also footnote 43 p. 22) (Osstyn 2010) explain that voice-over commentary, whether it is in English, German or French, must be translated rather freely. The script often has to be rewritten, which involves the cutting and restructuring of complex sentences to form a fluent Dutch text that is easy to comprehend as the audience hears the text only once. According to the guidelines, sound Dutch off-screen dubbing requires rather short sentences, where metaphors and flowery descriptions, which are typical of English commentary, are avoided, whereas Dutch is less emotional and more measured. As far as domain-specific terminology is concerned, the VRT asks to provide a clear translation for terms and adds that sometimes a description is less patronising than an unintelligible explanation.
The examples below show translational choices according to these guidelines.

Omission of the metaphor:

> *It's part of a southern river system that flows underground here, carving holes into the limestone* ***like a Swiss cheese.***
> *Het hoort bij een rivierenstelsel dat onder de grond loopt en gaten uit de kalksteen slijt.*
> *[Back translation: It's a part of the river system that flows underground and carves holes into the limestone]*
> *(The Earth Machine – Land, 2011)*

Omission of an adjective:

> *A giant fig, surprisingly and* ***persistently*** *green, wafts its thirsty roots across the ground.*
> *Een wurgvijg - hij ziet nog verrassend groen - zwaait met zijn dorstige wortels over de grond.*
> *[Back translation: A giant fig – he looks surprisingly green – swings with its thirsty roots over the ground.]*
> *(Madagascar, 2011)*

Restructuring of the sentence:

> *22nd of February 2011. The cathedral spire in Christchurch New Zealand falls, as an earthquake leaves 182 dead.*
> *22 februari 2011. Christchurch in Nieuw-Zeeland wordt getroffen door een zware aardbeving. De kathedraal van de stad stort bijna helemaal in. Er vallen 182 doden.*
> *[Back translation: 22 February 2011. Christchurch in New Zealand is hit by a heavy earthquake. The cathedral of the town falls down almost entirely. There are 182 dead.]*
> *(The Earth Machine – Land, 2011)*

Description of a term:

> *The Pacific Rim is an area of intense earthquake activity.*
> *De Pacific Rim,* ***de landen rondom de Stille Oceaan****, is een gebied waar erg veel aardbevingen voorkomen.*
> *[Back translation: The Pacific Rim, the countries around the Pacific Ocean; is an area with lots of earthquakes.]*
> *(The Earth Machine – Land, 2011)*

Obviously, the translator must also take into account the images (off-screen dubbing and visual image have to match perfectly). In addition, the translator is encouraged to add indications for the voice talent and the voice director (e.g. pauses or an alternative translation).

An example of an alternative translation choice, suggested by the translator for the voice director:

> *Planet Earth is unique. An engineering marvel. Driven by heat from within.*
> *De planeet Aarde is uniek. Een ingenieuze machine (****OF: Een wonder van technisch vernuft: al impliceert 'technisch' dat de mens erachter zit****), aangedreven door warmte die van binnenuit komt.*
> *[Back translation: The planet Earth is unique. An engineering machine (OR: A wonder of technical contrivance: even though 'technical' implies human influence), driven by heat from within.]*
> *(The Earth Machine – Land, 2011)*

The guidelines also mention that translations may be altered by the voice director before or during the recording of the voice in the studio, a practice that is quite common in dubbing, where the dubbing director and the voice talent may modify the text while recording (Chaume 2014).

In this case, the two different translation stages resulting from the interventions in the course of the translation process have spawned two translation versions for each production. In other words, the VRT database and the corpus under study contain two translations of each script: the first translation and the 'adapted' version (with the voice director's modifications). As a consequence, it had to be tested to what extent the differences between the two versions affected the translation of the domain-specific terminology, since the latter is a crucial factor in this dissertation.

A count of the terminological units was done in order to proceed with the most 'term-rich' version. The experiment corpus was used for this purpose as this consists of the texts used for the experiments.

All term units of the first translation and the adapted version were manually labelled (Chapter 3 provides methodological details concerning the labelling), counted, and compared with the total number of source text words and with the number of unique source text words, that is all the words excluding the repetitions. Term tokens means the total number of terms, repetitions included, term types means the total number of terms, repetitions excluded. Table 7 below shows the results of the comparison.

Table 7 The number of term tokens and types in the first translation compared to the adapted version

	First translation		Adapted version	
	% term tokens vs. tot. number of ST words	% term types vs. tot. number of unique ST words	% term tokens vs. tot. number of ST words	% term types vs. tot. number of unique ST words
Madagascar	3.1	8.3	3.1	8.3
The earth machine	15.7	30.3	15.4	29.9
The secret world of pain	7.9	15.3	7.7	14.9

The figures indicate negligible differences since the adaptation of the first translation related mainly to the indications for the voice talent, the sentence structure, repetitions, pronouns, either for the sake of text fluency or to synchronise the off-screen dubbing and the visual images. Some examples of adaptations made by the voice director in the studio are given in Table 8 below.

Table 8 Examples of adaptations made by the VRT voice director

English source text	First translation	*Back translation first translation*	Adapted version	*Back translation adapted version*
Around the world, there is a remarkable group of people who hold the clue to one of science's greatest mysteries: Pain. *(The secret world of pain, 2011)*	Vandaag maken we kennis met een groep mensen die de sluier van een eeuwenoud medisch mysterie lichten. **Wat is pijn**?	*Today we meet a group of people who will explain us an ancient medical mystery. What is pain?*	Vandaag maken we kennis met een groep mensen die de sluier van een eeuwenoud medisch mysterie lichten. **– P- Pijn** (the 'P' is an indication for the voice talent and means 'pause')	*Today we meet a group of people who will explain us an ancient medical mystery. Pain*

Totally cut off from the rest of the world, these castaways made this island their own, gradually evolving into a collection of wildlife that's strange, rare and utterly unique. *(Madagascar, 2011)*	Afgesneden van de rest van de wereld gingen ze hier hun eigen weg en gaandeweg ontwikkelde zich een zeldzame en unieke collectie dieren. **Het zijn echt wel buitenbeentjes**. Meer dan tachtig procent van de soorten hier zie je nergens anders op aarde.	*Cut off from the rest of the world, they went here their own way and gradual a collection of wildlife evolved into a rare and unique collection of animals. They are real outsiders. More than eighty percent of the species here, you never see anywhere else on earth.*	Afgesneden van de rest van de wereld gingen ze hier hun eigen weg en gaandeweg ontwikkelde zich een zeldzame en unieke collectie dieren. Meer dan tachtig procent van de soorten hier zie je nergens anders op aarde.	*Cut off from the rest of the world, they went here their own way and gradual a collection of wildlife evolved into a rare and unique collection of animals. More than eighty percent of the species here, you never see anywhere else on earth.*
Its unusual geological history, its isolation, and its resting place in the tropics, were to shape Madagascar's fortunes. *(Madagascar, 2011)*	**Het** woelige geologisch verleden en de geïsoleerde ligging pal in de tropen zouden de toekomst van het eiland bepalen.	*The bustling past and the isolated position in the tropics were to shape the island's future.*	**Zijn** woelige geologische verleden en zijn geïsoleerde ligging pal in de tropen zouden de toekomst van het oudste eiland op aarde bepalen.	*Its bustling past and the isolated position in the tropics were to shape the island's future.*

The number of domain-specific terms is slightly lower in the adapted version and other changes in the adapted version do not influence the domain-specific terminology. That is why the first translation was used for the creation of the gold standard, the testing of the automatic term extraction systems and for the creation of the automatic bilingual glossary.

The following chapter explains the methodology used to achieve these goals.

Chapter 3
RESEARCH METHODOLOGY

3.1 Introduction

The main research question of this dissertation is:
'Does the integration of an automatically generated and/or a manually generated bilingual glossary into the translation process reduce the translators' workload?'

In order to operationalise this over-arching question, it was subdivided into five, concrete research questions:

1) Does the corpus of nature and science documentaries of this dissertation contain domain-specific terminology?

2) Is the terminology used in off-screen dubbing for documentaries specific enough to be detected by automatic terminology extraction systems?

3) Does the integration of a domain-specific, automatically and manually generated bilingual glossary into the translation process reduce the process and pause time before terms (which are indicators of translation efficiency and speed) of Master's students in translation and professional translators?

4) Does the number of terminological errors decrease when translators work with a bilingual glossary?

5) To what degree do Master's students in translation and professional translators use the glossary? Do they consult it more as they become more used to having it at hand?

The methodology used to answer these questions will be explained in the following sections.

3.2 The preparatory stage

3.2.1 Sentence alignment

Section 1.2 ‘Insights from AVT studies’ concludes that under certain conditions of information treatment and communication, the lexical features in specialised discourse - including what is called in my corpus ‘science and nature’ documentaries - constitute the term units. In other words, under certain conditions, terminology is a linguistic feature of informative text types. This led me to hypothesise an affirmative answer on research question one.

However, in order to confirm whether or not the corpus contains term units, as defined in greater detail below, the domain-specific terminology of a representative sample of source texts, called ‘the preparatory corpus’ presented in Chapter 2 (see Table 5) was manually labelled. First, i.e. before labelling the terms, all source and target sentences of the off-screen dubbing were aligned.

The corpus alignment was carried out by means of a freeware translation memory software called Similis®[1]. This system contains a monolingual lexicon for each language processed and runs a linguistic analysis that sees sentences as a series of syntactical units called 'chunks'. The chunks are identified by algorithms that recognise grammatical categories of individual words and then group these categories of words into word groups e.g. verbal groups or nominal groups (Planas 2005). This linguistic analysis is the underlying system for the alignment[2]. The alignment was corrected manually in order to obtain a 100% match between source and target text before starting the terminology extraction.

3.2.2 Manual terminology extraction

In order to accurately answer research question one, a clear-cut definition of what a ‘domain-specific term’ is required. Section 1.3 ‘Insights from terminology studies’ already addressed some important definitions detailing what a ‘term’ is in the literature. Yet, the most relevant approaches for this study are offered by Wright (1997) and Bowker (2008) who highlight two different theoretical perspectives which I will apply in order to manually label the terms in the preparatory corpus.

[1] http://similis.org/linguaetmachina.www/index.php

[2] Retrieved from http://emmanuel.planas.iplv.fr/wp-content/uploads/2012/12/20101108_Cours_Similis_0_Introduction.pdf p.29

3.2.2.1 Definitions and criteria

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 (2008) 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." (p.286, my italics).

These definitions mention two important criteria (marked in italics) that are crucial when selecting terms in a text, i.e. when determining what is a 'term' and what is a 'word'. Firstly, while labelling manually, annotators must carefully consider the importance of a word in its domain-specific context, that is, the frequency of its occurrence in the domain at hand versus its occurrence in other domains and versus its occurrence in general language. Secondly, annotators must consider both single words and multi-word units. These criteria are defined by Kageura and Umino (1996) as termhood and unithood and discussed in Section 1.3.4. While termhood "refers to the degree that a linguistic unit is related to [...] domain-specific concepts", unithood "refers to the degree of strength or stability of syntagmatic combinations or collocations" (Kageura and Umino 1996, p. 260-261). Some of the statistical filters developed to measure termhood and unithood (C-Value, TF IDF and Log-likelihood Ratio, mentioned in Section 1.3) will be covered more extensively in Chapter 4 'Experiments'.

A third criterion, relevant for bilingual terminology extraction, is the quality of the translation. Macken, Lefever and Hoste (2013) note that a bilingual glossary for use by translators must consist of valid translation pairs. As the corpus for this study was provided by the Flemish public broadcaster, the translations are considered to be of outstanding quality.

3.2.2.2 Manual labelling of the preparatory corpus and selection of the experimental corpus

For the manual labelling of the domain-specific terminology in the preparatory corpus, one annotator labelled all the terms of the ten aligned texts on the basis of the criteria mentioned above and ranked them into three sub-categories, based on the degree of relatedness to the domain-specific context. Sub-category one stands for very strongly related to the context, two for strongly related and three for not so strongly related but still related. The examples below illustrate how the annotation is done. Appendix 1 explains the criteria and guidelines for manual terminology extraction provided to the annotators. These annotation guidelines were those used for manual term extraction in the TExSIS project[3].

[3] https://www.lt3.ugent.be/projects/texsis

The domain-specific terms are all 1:1 translations. The number 1 refers to sub-category 1.

1) As well as being Madagascar's equivalent of hedgehogs, tenrecs also take the place that moles and shrews would occupy anywhere else in the world.
Tenreks zijn niet alleen het plaatselijke equivalent van onze egel, ze nemen ook de plaats in van onze mollen en spitsmuizen.
hedgehogs#egel#1
tenrecs#tenreks#1
moles#mollen#1
shrews#spitsmuizen#1
(*Madagascar,* 2011)

The subordinate clause in the first source text sentence, 'the world's tiniest reptile', is translated in the second target text sentence. Consequently, the term 'reptile' has no correspondent term in the first Dutch target sentence and vice versa 'reptiel' in the second target sentence has no correspondence in its source text, which is indicated by three dots: '...'. The number 2 refers to sub-category 2.

2) A pygmy chameleon, the world's tiniest reptile, tiptoes through the leaf litter on the steep volcanic slopes.
Een dwergkameleon kruipt voetje voor voetje over de rottende bladeren de helling op.
pygmy chameleon#dwergkameleon#1
reptile#...#2
She's so tiny, she's scarcely bigger than an ant.
Dit is het kleinste reptiel op aarde. Ze is amper groter dan een mier.
ant#mier#1
...#reptiel#2
(*Madagascar,* 2011)

This example shows a multi-word term. Multi-word terms must be a syntactical unit, called phrase, and are broken up until the base word. The number 3 refers to sub-category 3.

3) A virtual planet earth.
Een virtuele planeet Aarde.
virtual planet earth#virtuele planeet Aarde#1
virtual#virtuele#3
planet earth#planeet Aarde#2
planet#planeet#3
earth#Aarde#3
(*Earth Machine – Land,* 2011)

Example No. 4 illustrates how locations are indicated. 'NE' stands for 'named entity'. Organisations are indicated as 'NE_ORG' (see example No. 5), other names different from locations, organisations etc. are miscellaneous: 'NE_MISC'.

4) To create mountain ranges like the Alps.
En zo ontstaan bergketens, zoals de Alpen.
mountain ranges#bergketens#3
Alps#Alpen#NE_LOC
Deep enough to accommodate the Empire State Building.
Zo diep zelfs dat de Eiffeltoren er bijna anderhalve keer in kan.
Empire State Building#Eiffeltoren#NE_MISC
(*Earth Machine – Land*, 2011)

This last example shows that the terms are not lemmatised. Declensions, conjugations, plural forms are maintained as they feature in the source and target text. The three asterisks indicate non-terms between two parts of a term unit.

5) Now in Seattle at the Harbour View Medical Center, scientists are harnessing this new understanding of the brain to help burn victims who suffer excruciating pain on a daily basis.
Onze volgende halte is het brandwondencentrum van Seattle. De patiënten hier lijden soms ondraaglijke pijnen. En ze hebben een revolutionaire manier ontwikkeld om hen te helpen.
Harbour View Medical Center#brandwondencentrum#NE_ORG
*suffer ***pain#lijden*** pijnen#1*
suffer#lijden#2
pain#pijnen#1
(*The Secret World of Pain,* 2011)

Once all the term units were listed, the term types i.e. the unique terms, were counted and compared to the number of unique words in the source text. A computer script was written to count the number of unique words. Table 9 below shows the results in absolute numbers and in percentages for each documentary.

Table 9 The number of term types compared to the number of unique words in the source text in percentages

Subject	Title	Total number of term types	Total number of unique words in the source text	% term types vs unique words in the source text
Wildlife	Around the world in 80 gardens-Australia	75	1020	7.4
Wildlife	Big cat diary 7	24	313	7.7
Wildlife	Madagascar	92	1095	8.4
Wildlife	Polar bear	189	998	18.9
Wildlife	The empire of the ants	176	1109	13.4
Earth and space	How earth made us	221	1088	20.3
Earth and space	The earth machine-Land	337	989	34.0
Earth and space	The magnetic pole	325	1111	29.3
Human body	E-numbers	185	785	23.6
Human body	The secret world of pain	168	751	22.4

The figures range from 7.4% to 34.0% with an average of 18.8%. This means that on average, almost one out of five unique words is a term. Yet, documentary texts are written and translated for a heterogeneous target audience, as discussed in Section 1.2.3. Consequently, the term units labelled in the corpus vary from very domain-specific (sub-category 1) to little domain-specific (sub-category 3). Table 10 below illustrates the sub-categories compared to the number of unique words in the source text.

Table 10 The term types in sub-categories compared to the number of unique words

Subject	Title	Channel	Sub-cat.1	Sub-cat.2	Sub-cat.3
Wildlife	Around the world in 80 gardens-Australia	Eén	1.0	1.9	4.5
Wildlife	Big cat diary 7	Eén	1.9	2.6	3.2
Wildlife	Madagascar	Canvas	5.4	2.2	0.7
Wildlife	Polar bear	Canvas	7.3	9.2	2.1
Wildlife	The empire of the ants	Canvas	6.2	1.2	6.0
Earth and space	How earth made us	Canvas	10.8	7.0	2.5
Earth and space	The earth machine-Land	Canvas	9.4	8.0	16.6
Earth and space	The magnetic pole	Canvas	6.9	10.4	11.9
Human body	E-numbers	Canvas	3.2	8.7	11.7
Human body	The secret world of pain	Canvas	13.6	3.3	5.5

The two documentaries broadcast by channel Eén (*Around the World in 80 Gardens* and *Big Cat Diary*) contain little domain-specific terminology in general and few terms strictly related to the domain. As was explained in Chapter 1 'Insights from documentary studies', Eén aims to reach a broad target audience and entertainment rather than education and information. This may explain why the documentaries broadcast by Eén contain less domain-specific terminology than those broadcast by the cultural channel Canvas.

In other words, the above shows that documentaries meant for educational and informative purposes contain domain-specific terminology, which provides a positive answer to research question one. Moreover, these terms can be very specialised, for example 'giraffe necked weevil' *(Madagascar)*, 'mantle plume' *(The Earth Machine)* or 'trans-cranial magnetic stimulation' *(The Secret World of Pain).* Audiovisual documentary translators, who are typically specialised in a translation mode, not in a field or a topic as explained in Section 1.2.1 have to deal with this specialised terminology in many different subjects.

The hypothesis is that a bilingual glossary constitutes an additional resource of specialised information that gives added support to audiovisual translators. The degree to which a glossary provides support (research question 3, 4 and 5) will be investigated by means of experiments discussed in Chapter 4. Research question 2 is considered in the next section.

3.2.2.3 Creation of the gold standard

Research question 2 is the question whether or not the terminology in documentaries is specific enough to be detected by automatic terminology extraction systems. In order to investigate that question, an exhaustive and accurate bilingual glossary, called gold standard (GS), needed to be created manually as an objective means for testing the automatic term extraction systems.

Three annotators labelled the domain-specific terminology of the experimental corpus manually according to the same annotation guidelines as those used for the preparatory corpus and specified in Appendix 1. Next, 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)

F-scores are calculated by considering the annotations of one annotator as the GS, and measuring precision and recall of the second annotator on that gold standard set of annotations. Precision gives an indication of whether or not the proposed terms are relevant, whereas recall measures the capability to retrieve *all* relevant terms in a given text. In terminology extraction research, it is common practice to calculate only precision. Yet, as a GS was created, it enabled me to also calculate recall and to have a clear view on the number of terms in the texts and the coverage of the terminology extraction. Table 11 lists the inter-annotator figures of the manual term extraction.

Table 11 Rate of agreement expressed in precision, recall and F-score between the 3 annotators

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

The rates indicate that e.g. for *Madagascar,* annotator 1 labelled only 24.86% of the terms labelled by annotator 2 (precision) whereas 71.32% of all terms labelled by annotator 2 were also labelled by annotator 1 (recall). The output suggests a strong disagreement between annotator 1 and annotator 2 and 3 for *Madagascar* and in general, a fuzzy boundary between what is to be considered a 'term' and a 'non-term'. In *Madagascar*, for instance, the domain of terminology can be restricted to animals and plants, but it could also be extended to encompass all natural phenomena including rock, for instance. In *The Earth Machine*, terms could 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, a form of subjectivity that can be reduced by providing very specific instructions, even though the decision on whether or not to classify a word or a group of words as a terminological unit will inevitably remain a personal choice to some extent. However, for translational purposes, a broad interpretation of what a term is, is more useful. Translators might also want to check the bilingual glossary in case of doubt about the spelling of a term, to verify the accuracy of a translation or they might simply want to use the glossary as a source of inspiration. For this reason, the GS[4] was created, assembling the terms of all three annotators into one glossary, without any filtering. By doing so, a slightly fuzzier glossary was obtained. Because of the hybrid text type containing both technical and creative elements, and aimed at a mixed audience, glossaries may contain 'not really technical' terms, in addition to the strictly domain-specific terms (see also the discussion in Section 1.2.1 on the specific challenges of translating documentaries).

3.2.3 Automatic terminology extraction

By means of this GS, the accuracy of automatic terminology extraction systems with off-screen dubbing for documentaries could now be evaluated. Two commercial systems for bilingual term extraction (SDL Multiterm Extract 2011 Trados® and Similis®) and one system (TExSIS) developed by the University of Ghent[5] (Macken et al., 2013) were tested. The underlying technology of these systems was explained in detail in Section 1.3.4.

As the terminology lists resulting from the bilingual term extraction (hereinafter referred to as glossaries) of the three systems were compared to the terminology lists of the GS, it was possible to calculate the rate of agreement expressed in terms of precision, recall and F-score in Table 12 below.

[4] The complete GS glossaries as they were used for the experiments, are included in Appendix 2.

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

Table 12 Rate of agreement for the bilingual terminology extraction of the GS and the three automatic term extraction systems

Agreement scores between systems in %		Madagascar	The Earth Machine	The Secret World of Pain
GS vs Similis	Precision	33.33	28.78	24.32
	Recall	3.77	7.53	6.36
	F-score	**6.77**	**11.94**	**10.08**
GS vs SDL	Precision	32.08	25.37	21.43
	Recall	7.11	9.60	5.30
	F-score	**11.64**	**13.93**	**8.50**
GS vs TExSIS	Precision	55.97	58.67	53.66
	Recall	15.69	16.57	7.77
	F-score	**24.51**	**25.84**	**13.58**

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 correspond exactly to the terms labelled manually in the GS. The higher the recall score, the better the system is capable to retrieve all relevant terms, in other words the more complete the glossary is compared to the GS.

The precision scores of TExSIS[6] are better than those of Similis and SDL for all three texts. By way of comparison, in the tests Macken et al. (2013) conducted on technical texts of the automotive industry, the TExSIS precision scores were 61.95 to 66.55 whereas the present results vary from 53.66 to 58.67.

Yet, both recall and F-score are low for all systems. This can be explained by the audiovisual nature of documentary texts. The VRT guidelines say 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.

The excerpts in Table 13 below, from *The Secret World of Pain,* illustrates the free Dutch translation.

[6] The complete TExSIS glossaries as they were used for the experiments, are included in Appendix 2.

Table 13 Excerpts from *The Secret World of Pain,* illustrating the free Dutch translation

English commentary	Dutch off-screen dubbing	Back translation
Developing radical new therapies targeting the source of pain, together this could transform the lives of millions.	Ze richten hun therapieën op de plaats waar pijn ontstaat.	They target their therapies to the place where pain starts.
Deep in the heart of the Tuscan countryside live a family who could help unlock one of the great questions about pain: why we feel it so differently. The Marsillis.	Waarom ervaart iedereen pijn anders? Voor het antwoord op die vraag trekken we naar het prachtige Toscaanse platteland. Want daar woont een wel heel bijzondere familie. Dit zijn de Marsilli's.	Why does everybody experience pain differently? For the answer to this question we go the beautiful Tuscan countryside. Because there lives a very special family. These are the Marsillis.

3.3 The experiments

The answers to research questions 3, 4 and 5 were obtained by means of translation experiments.

First, a pilot study with Master's students in translation was conducted, after which an experiment with professional translators was carried out. The experimental set-up and the analysis of the results will be addressed in Chapter 4.

Research question 3 tackles the central issue of this study. It aims to find out whether or not the integration of a domain-specific, automatically and manually generated bilingual glossary into the translation process reduces the process and pause time before terms (which are indicators of translation efficiency and speed) of Master's students in translation and professional translators.

As "writing fluency and flow reveal traces of underlying cognitive processes" (Leijten & Van Waes 2013, p. 360) and pauses are indicators of cognitive effort in all types of language production (Lacruz et al., 2014), the writing process needs to be observed. Keystroke logging programmes generally log and time stamp activity in order to

reconstruct and describe text production processes. Inputlog[7], a logging software developed at the University of Antwerp, was chosen for the current research because of its compatibility with Windows environments and its writing-oriented design which is entirely unobtrusive (Leijten & Van Waes, 2013).

While the participants were translating, Inputlog registered their entire translation process. Specific features of the software allow the researcher to analyse two crucial aspects indicating whether bilingual glossaries reduce the translator's workload viz. total translation time and the pause time before terms[8].

Similarly, in post-editing research, Kring (2001) identifies the total time spent to perform the task as 'temporal effort' and the errors and the necessary steps for correction as 'cognitive effort'. The latter has been implemented by Lacruz et al. (2014) who related cognitive effort to the pause-to-word ratio (the number of pauses per word in a post-edited MT segment). In the present dissertation, the cognitive effort is represented by the total pause time before each term. Yet, as 'temporal and cognitive effort' are linked to post-editing research, the indicators used in the thesis under study are called 'total process time' and 'pause time before terms'.

Inputlog also allows analysing the search for information, recording every keystroke and mouse movement and thus every webpage, server-based dictionary or locally uploaded file the participant consults. This information provides insights to answer research question five (to what degree do the participants use the glossaries?).

The features used for this study were the 'record' and the 'analyse' module. The record module logged all keystroke and mouse data in the Microsoft Word page used for the translation, together with a time stamp in milliseconds, while the analyse module produced the general logging file and the summary analysis needed to identify the total process time and the pause time before terms, as these two data were crucial to determine the efficacy of a bilingual glossary.

The total process time of the translations was provided in milliseconds by the summary analysis generated in Inputlog. The pause time before terms could be deducted from the general analysis Inputlog supplies, which 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 of the general analysis represent every log event, the cursor position, the document length, the start and end time of every event in milliseconds, and can therefore be used to calculate the action and pause times. In order to determine the pause times before the terms[9] needed for the experiments, the log files were scrolled manually, selecting the rows belonging clearly to a pause time before terms, i.e. the event from the

[7] http://www.inputlog.net/ (accessed 20 October 2014).

[8] This is the time the participant spent to look up the information needed to translate a term.

[9] 'Terms' means 'labelled as such' in the manually extracted glossary about which is reported above.

moment the participant entered a dictionary, the internet or the bilingual glossary until he/she returned to the Wordlog document in which the translation was written down. If the participant surfed 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.

Figure 1 shows an example of the Inputlog general analysis output, illustrating the output features. Every row represents one log event (second column 'id'). The first row of this example indicates that the participant entered the TExSIS-glossary, then used the CRTL-F function to look up the term 'mass' (vertical digits, row 1418 – 1421). The cursor position and the document length are represented ('positionFull' and 'doclengthFull'), as well as the number of characters produced (charProduction). The next columns show the start time and the end time of every event in milliseconds. These data are used to calculate the action and pause time, in this case the sum of the milliseconds indicated in row 1413 up to row 1429. An algorithm identifies the pause location and provides a classification ('pauseLocation' and 'pauseLocationFull') and in the last columns the mouse clicks are represented by x/y values.

id	type	output	posit	docl	char	startTim	startClock	endTim	endClock	actio	pauseTir	pause	pauseLocat	x	y
1413	focus	Microsoft Excel - Hunt for the Higgs_TXtermenlijst	2636	5469	6107	2209828	00:36:49.82	2209828	00:36:49.82	0	0	10	TRANSITION		
1414	keyboard	LCTRL	2636	5469	6107	2209828	00:36:49.82	2210125	00:36:50.12	297	0	2	BEFORE WORDS		
1416	keyboard	LCTRL + F	2636	5469	6107	2210047	00:36:50.04	2210141	00:36:50.14	282	31	13	UNKNOWN		
1417	focus	Find and Replace	2636	5469	6107	2210125	00:36:50.12	2210125	00:36:50.12	0	0	10	TRANSITION		
1418	keyboard	m	2636	5469	6107	2210391	00:36:50.39	2210484	00:36:50.48	93	0	1	WITHIN WORDS		
1419	keyboard	a	2636	5469	6107	2210422	00:36:50.42	2210516	00:36:50.51	94	31	1	WITHIN WORDS		
1420	keyboard	s	2636	5469	6107	2210625	00:36:50.62	2210703	00:36:50.70	78	203	1	WITHIN WORDS		
1421	keyboard	s	2636	5469	6107	2210766	00:36:50.76	2210875	00:36:50.87	109	141	5	AFTER SENTENCES		
1422	keyboard	RETURN	2636	5469	6107	2212219	00:36:52.21	2212281	00:36:52.28	62	1453	4	BEFORE SENTENCES		
1423	mouse	Movement	2636	5469	6107	2212969	00:36:52.96	2213922	00:36:53.92	953	750	13	UNKNOWN	624	358
1424	mouse	LEFT Click	2636	5469	6107	2214016	00:36:54.01	2214078	00:36:54.07	62	94	2	BEFORE WOR	624	358
1425	focus	Microsoft Excel - Hunt for the Higgs_TXtermenlijst	2636	5469	6107	2214094	00:36:54.09	2214094	00:36:54.09	0	0	10	TRANSITION		
1426	mouse	Movement	2636	5469	6107	2214094	00:36:54.09	2214625	00:36:54.62	531	0	13	UNKNOWN	932	792
1427	focus	TASKBAR	2636	5469	6107	2214719	00:36:54.71	2214719	00:36:54.71	0	0	10	TRANSITION		
1428	mouse	LEFT Click	2636	5469	6107	2214719	00:36:54.71	2214797	00:36:54.79	78	0	2	BEFORE WOR	932	792
1429	focus	Microsoft Excel - Hunt for the Higgs_TXtermenlijst	2636	5469	6107	2214797	00:36:54.79	2214797	00:36:54.79	0	0	10	TRANSITION		
1828	focus	Large Hadron Collider - Wikipedia - Mozilla Firefox	3784	5652	6450	2299609	00:38:19.60	2299609	00:38:19.60	0	0	10	TRANSITION		
1829	mouse	Movement	3784	5652	6450	2299609	00:38:19.60	2300484	00:38:20.48	875	0	13	UNKNOWN	1028	53
1830	mouse	LEFT Click	3784	5652	6450	2300641	00:38:20.64	2300844	00:38:20.84	203	157	2	BEFORE WOR	1028	53
1831	mouse	Movement	3784	5652	6450	2300688	00:38:20.68	2300813	00:38:20.81	125	0	13	UNKNOWN	544	47
1832	mouse	Movement	3784	5652	6450	2300859	00:38:20.85	2300859	00:38:20.85	0	46	13	UNKNOWN	544	46
1833	keyboard	p	3784	5652	6450	2301375	00:38:21.37	2301453	00:38:21.45	78	516	1	WITHIN WORDS		
1834	keyboard	h	3784	5652	6450	2301609	00:38:21.60	2301750	00:38:21.75	141	234	1	WITHIN WORDS		
1835	keyboard	o	3784	5652	6450	2301688	00:38:21.68	2301813	00:38:21.81	125	79	1	WITHIN WORDS		
1836	keyboard	t	3784	5652	6450	2301813	00:38:21.81	2301906	00:38:21.90	93	125	1	WITHIN WORDS		
1837	keyboard	o	3784	5652	6450	2301875	00:38:21.87	2302047	00:38:22.04	172	62	1	WITHIN WORDS		
1838	keyboard	n	3784	5652	6450	2301969	00:38:21.96	2302109	00:38:22.10	140	94	5	AFTER SENTENCES		
1839	keyboard	RETURN	3784	5652	6450	2302453	00:38:22.45	2302547	00:38:22.54	94	484	4	BEFORE SENTENCES		
1840	mouse	Movement	3784	5652	6450	2303016	00:38:23.01	2304500	00:38:24.50	1484	563	13	UNKNOWN	448	319
1841	focus	photon - Google zoeken - Mozilla Firefox	3784	5652	6450	2303078	00:38:23.07	2303078	00:38:23.07	0	0	10	TRANSITION		

Figure 1 An example of Inputlog's general analysis

Figure 2 below is an excerpt of the summary analysis Inputlog provides. For the present study, the general 'Total Process Time' was used in order to compare the total process time of the participants. The Inputlog data concerning the process and pause time were statistically analysed with parametrical tests, as will be explained in detail in Chapter 4 when the experiments will be discussed.

inputlog

Summary Logging File

Meta Information	
Session Identification	
Parameters	
Pause Threshold (ms)	500
Process Information	
Product Information	
Product/Process	
Process Time	
General	0:25:24
Total Process Time (s)	1,524.860
Number of P-Bursts	411
Mean Process Time P-Bursts (s)	3.710
Standard Deviation P-Bursts	2.807
Total Pause Time	0:10:58
Total Pause Time (s)	658.513
Number of Pauses	411
Mean Pause Time (s)	1.602
Standard Deviation Pause Time	2.078
Active Writing Time	0:14:26
Total Writing Time (s)	866.347

Figure 2 An example of Inputlog's summary analysis featuring the general 'Total Process Time'

A subsequent issue that leads to understand the translators' workload, namely working with and without a bilingual glossary, is addressed in research question 4. This question aims to find out whether or not the number of terminological errors decreases when translators work with a bilingual glossary. For this purpose, a manual calculation of all terminological errors was carried out. A 'terminological error' is considered to be a term from the GS that was translated with a term not corresponding to the original target text, the translation made by the VRT translators. Appendix 5 – uploaded on my website[10] because of the large and numerous files - provides detailed information about the terminological errors made by the students and professionals. After having completed the count, the results were then used for parametrical tests and bar graphs which compare the number of errors that the participants made with and without glossary in percentages (see Chapter 4).

[10] https://www.uantwerpen.be/nl/personeel/sabien-hanoulle/mijn-website/

To turn to research question 5, 'To what degree do Master's students in translation and professional translators use the glossary? Do they consult it more as they become more used to it?' This was investigated on the basis of the "before term" pause time extracted from Inputlog's general analysis. By scrolling the log files manually, the rows could be selected in which the participants entered a dictionary, the internet or the bilingual glossary until they returned the Wordlog document to write down the translation. The pause times for internet searches were then added to the pause times for a dictionary search. The absolute and average pause times of the searches on the internet and in dictionaries were then compared with the pause times for searches in the bilingual glossary by means of a bar graph. The evolution of the use of the sources was represented in a line graph (see Chapter 4).

The following chapter deals with the students' and with the professional translators' experiment.

Chapter 4
TRANSLATION EXPERIMENTS WITH BILINGUAL GLOSSARIES

4.1 Introduction

In order to proceed with the translation experiments and to determine the impact of domain-specific, bilingual glossaries on translators' workload and workflow, a manually and an automatically extracted bilingual glossary of each documentary had to be drawn up.

Three documentaries were selected as experimental corpus, viz. *Madagascar – Island of Marvels* (2011), *The Earth Machine – Land* (2011) and *The Secret World of Pain* (2011). The latter was selected, because it had, by far, the highest number of terms in sub-category 1 for the subject 'human body' (see Chapter 3, Table 10) and consequently, the highest degree of specialisation of terms. For the subjects 'wildlife' and 'earth and space', the difference between the number 1 and the number 2 in sub-category 1 was not that marked. For this reason, I verified manually which were the most problematic terms for a translator i.e. which would require the most research. Frequency of domain-specific terminology was not taken into consideration as the most specialised terms only occurred once or twice (e.g. giraffe necked weevil, Grandidier's vontsira, giant mongoose, igneous, free-floating blocks of concrete). *Madagascar* and *The Earth Machine - Land* were selected.

The manual glossaries corresponded with the GS explained in Section 3.2.2.3. For the automatically extracted glossaries, the termhood values were calculated. As explained in Section 1.3.4., there are several statistical filters to measure termhood, all based on frequency and length: C-Value (Frantzi et al., 2000), TF IDF (Spela Vintar 2004) and Log-Likelihood ratio (Dunning 1993), amongst others. A formula and brief description of the first two is given below, while the last is illustrated with an example.

The **C-value** is a measure designed to recognise and extract multi-word terms (Frantzi et al., 2000). After a shallow linguistic analysis (involving PoS tagging and chunking), they

focus on statistical filtering taking into account the total frequency of occurence of the candidate string in the corpus, the frequency of the candidate string as part of other longer candidate terms, the number of these longer candidate terms and the length of the candidate string. C-values above a predefined threshold are considered terms. Frantzi et al. (2000) provide the following formula:

$$C - Value(a) = (length(a) - 1)(freq(a) - \frac{t(a)}{c(a)}$$

a = collocation

t(a) = frequency of *a* in longer candidates of collocations

c(a) = number of longer candidates of collocations including *a*

In 2008, an adaption was made to this formula by Vu, Aw & Zhang (2008), who also introduced a unithood feature in the original formula.

The **TF IDF filter** (term frequency / inverse document frequency) is a way to score the importance of terms in a document based on how frequently they appear across multiple documents (Vintar 2004):

$$tf.idf(i,j) = \begin{cases} 1 + \log(tf_{i,j}))log\,\frac{N}{df_i}, tf_{i,j} \geq 1 \\ 0, tf = 0 \end{cases}$$

where *N* is the number of all documents in a corpus, tf_{ij} is the frequency of the term w_i in the document corpus d_i and the document frequency df_i is the number of documents where the term occurs at least once. As the C-value, the TF IDF filter applies a predefined threshold to distinguish words from terms.

The **Log-Likelihood Ratio** (LLR) is a statistical measure used to calculate terminologically relevant single-word terms and collocations, considering word frequencies weighted over two different corpora (Rayson & Garside, 2000). Words having a much higher or lower frequency than expected are assigned high log-likelihood (LL) values. LL values above a predefined threshold are considered terms, just as the former two termhood filters (Macken, Lefever, & Hoste, 2013).

In order to calculate LL, a frequency list is made for each corpus (the domain-specific and a background corpus). Then, LL is calculated for each word in the frequency lists. This is done by constructing a contingency table as is shown in Table 14, where c represents the number of words in the first corpus, while d corresponds to the number of words in the second corpus. Values a and b are called the observed values (O).

Table 14 Contingency table to calculate LL

	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	d	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 (Rayson & Garside, 2000):

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

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 LL:

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

By way of illustration, the LLR of the word 'crust' in *The earth machine* (2011) will be calculated manually.

Table 15 Contingency table to calculate LL of 'crust'

crust	Experimental corpus	General corpus[1]	Total
	23 (a)	5 (b)	28
	3564 (c)	3567418 (d)	3570982

[1] Invented figures

$E_1 = 3564 * 28/3570982 = 0.0279$

$E_2 = 3567418 * 28/3570982 = 27.97$

$$LL = 2 * (23 * ln(\frac{23}{0.0279})) + (5 * ln(\frac{5}{27.97}))$$

$$= 823.03 \qquad = 0.17$$

$$\downarrow \qquad \downarrow$$

$$(23 * 6.71) + (5 * -1.72)$$

$$(154.33) + (-8.6)$$

$$= 2 * 145.73$$

$$LL = 291.46^2$$

Table 16 below presents the results for the automatic term extraction of 'crust', with a slightly different LLR from the manual calculation due to the rounding off (291.54915).

Yet, it was not possible to predefine a threshold for the LLR in the automatically extracted glossaries of the present corpus. As already mentioned in Section 3.2.3., the recall figures for my glossaries, indicating the system's capability to retrieve all relevant terms, are low due to the small corpus. For this reason, an automatic system fed with linguistic and statistical filters (I refer to Section 1.3.4. for the strategies of automatic term extraction systems) not always recognises the corresponding target term. In the source and target sentences below (assigned only a 70% match by the alignment tool) the POS patterns (a linguistic filter) are different so no syntactic similarity is found. 'smoke-shrouded' is a compound adjective whereas its translation 'rook' (smoke) is a noun and 'rim of the volcano' has not been translated at all. In fact, the term smoke-shrouded rim is not recognised by the system, as illustrated in Table 16.

> The team must descend from the **smoke-shrouded rim** of the volcano into the searing heat of the crater.
>
> Hun klimwerk wordt bemoeilijkt door de **dichte rook** en de verzengende hitte die uit de krater komt.
> *[Back translation: Their climbing is being hindered by the dense smoke and the searing heat coming out of the crater.]*

A second problem for automatic term extraction in the present corpus is the low frequency of some domain-specific terms. Statistical filters like LLR or TF IDF are based on the frequency of a candidate string in a given corpus and its frequency in general texts, or look at the frequency of the candidate string as part of other longer candidate terms, the number of these longer candidate terms and the length of the candidate string (C-value). However, in my corpus of documentaries, some terms occurred only once or twice (see examples above) and were therefore not extracted by the system. Two other

[2] http://ucrel.lancs.ac.uk/llwizard.html calculates the LL of above figures at 291.58

examples can be seen in Table 16. Although their English source text and translation show that both terms were translated literally, the system failed to detect them.

'Super-continent' only features once, while the translation is a 100% match with the source sentence:

> In two hundred and fifty million years there will be no more America, Russia or China, but one giant **super-continent**.
>
> Over 250 miljoen jaar zal er geen sprake meer zijn van Amerika, of Rusland of China. Dan zal er maar 'n groot **supercontinent** meer zijn.

'Celcius' features twice; one of the translations is again a 100% match with the source sentence, while the other one has omitted 'celcius' in the translation and has a 67% match:

> *The temperature on the lakes surface is nearly 800 degrees **celcius**.*
>
> *De temperatuur aan de oppervlakte van het meer loopt op tot 800 graden **celsius**.*
>
> *The mantle is over 1,800 miles deep – and up to a blistering 2200 degrees **Celsius**.*
>
> *De mantel is ongeveer 2900 kilometer dik en de temperatuur loopt er op tot 2200 graden.*
>
> *[Back translation: The mantle is about 1,800 miles thick and the temperature rises up to 2200 degrees.]*

As a result, the automatically extracted glossaries serving translational purposes, as is the case in this dissertation, were composed of *all* the term *pairs* the system extracted - if the target term was missing, the source term was discarded as well - irrespective of the termhood values. Consequently, the automatic glossaries, generated by TExSIS for this dissertation, also contain some errors: e.g. 'anything – bestaan'[3] (Hunt for the Higgs glossary p. 153) or 'limestone – bodem'[4] (Madagascar glossary p. 117). Some terms offer an apparently wrong corresponding term e.g. 'ground – rotsen (rocks) and aarde (earth)'[5] (The earth machine glossary p. 132) or 'blip – afwijking'[6] (Hunt for the Higgs glossary p. 153) while others are incomplete e.g. 'pain – pijnmechanisme, pijnsysteem'[7] (The secret

[3] Source text: Without gaining mass, particles would have continued to fly through the universe at the speed of light, never clumping together to form you, me, blackboards – well, anything.
Target text: Stel dat deeltjes geen massa kregen. Dan zouden ze door het heelal blijven vliegen met de snelheid van het licht. Nooit zouden deeltjes samenklonteren om wat dan ook te doen ontstaan dat een vorm heeft: de aarde, mensen, voorwerpen... Niets ervan zou bestaan.

[4] Source text: For millions of years, this landscape was drowned, and layers of limestone formed underwater.
Target text: Dit stuk van het eiland lag miljoenen jaren lang onder de zeespiegel. Op de bodem vormden zich dikke lagen kalk.

[5] Source text: The ground can give way at any moment.
Target text: De rotsen kunnen op elk moment loskomen.

[6] Source text: The only way scientists can tell if a Higgs boson was there or not, is by looking for a statistical anomaly, some blip in the measurements that they can't otherwise account for.
Target text: Er is maar één manier om uit te maken of er een higgsdeeltje is geweest of niet: zoeken naar een statistische ongerijmdheid, een vreemde afwijking in de metingen waar geen andere uitleg voor te vinden is.

[7] Source text: These will ultimately shape how our pain system is wired up.
Target tekst: En het zijn net die dingen die ons pijnmechanisme beïnvloeden.

world of pain glossary p. 146) or 'giraffe – giraffenkever'[8] (Madagascar glossary p. 117). These partial translations have not been filtered out applying the freqRAT[9] filter for the same reason why no threshold has been predefined for the LLR: every suggestion can inspire the translator to find a creative solution, which is important for off-screen dubbing, as mentioned in Section 1.1.1 dealing with the genesis of documentaries and in Section 1.2.1 concerning the specific challenges of translating documentaries. The off-screen dubbing needs to avoid repetition and to match with the images, so finding synonyms or other creative solutions is important to keep the audiences' attention. It is this aspect, the combination of creative and technical issues, that makes off-screen dubbing and consequently, the terminology extraction for off-screen dubbing a particular challenge.

An excerpt of the automatically extracted glossary of *The Earth Machine – Land* is shown below in Table 16 (the column with the number of words, the number of syllables and the lemma is omitted due to lack of space). The output is generated by TExSIS. The complete glossaries as they were used for the experiments, are included in Appendix 2.

Table 16 An excerpt of the bilingual glossary of 'The earth machine – Land', as it was extracted by the system, from term 1 to term 60

words	words	< POS>	< mapped POS>	avgSyl	Termhood	CValue	Freq	LLR
		<	<					291.54915
crust	aardkorst	NN>	N>	1.00	3.518.884	0.03144	23	
		<	<					291.54915
crust	korst	NN>	N>	1.00	3.518.884	0.03144	23	
		<	<					354.12590
earth	aarde	NN>	N>	1.00	3.009.679	0.05393	40	
		<	<					354.12590
earth	aardkorst	NN>	N>	1.00	3.009.679	0.05393	40	
		<	<					354.12590
earth	aardoppervlak	NN>	N>	1.00	3.009.679	0.05393	40	
super-continent		< NN>	< N>	5.00	2.804.639	0.00144	1	37.67210
		<	<					202.28059
lava	lava	NN>	N>	2.00	2.779.389	0.02134	16	

Source tekst: For John, the results could uncover an important gene that's involved in pain mechanisms.
Target tekst: John is op zoek naar een belangrijk gen dat het pijnmechanisme stuurt.

[8] Source text: It's a giraffe necked weevil, and this is a male.
Target text: Dit is een giraffenkever. Een mannetje.

[9] A measure which compares the frequencies of the source and target terms extracted by TExSIS, meant to assess translation validity.

		<	<					202.28059
lava	vulkaan	NN>	N>	2.00	2.779.389	0.02134	16	
		<	<					123.76177
mantle	mantel	NN>	N>	1.00	2.197.002	0.01442	11	
		<	<					123.76177
mantle	buitenmantel	NN>	N>	1.00	2.197.002	0.01442	11	
		<	<					169.20295
planet	aarde	NN>	N>	2.00	2.163.439	0.02660	20	
		<	<					169.20295
planet	planeet	NN>	N>	2.00	2.163.439	0.02660	20	
		<	<					169.20295
planet	oceanen	NN>	N>	2.00	2.163.439	0.02660	20	
		<	<					118.39083
tectonic		NN>	N>	3.00	2.143.032	0.00901	8	
		<	<					158.38904
surface	aardoppervlak	NN>	N>	2.00	2.099.709	0.03244	24	
		<	<					158.38904
surface	grond	NN>	N>	2.00	2.099.709	0.03244	24	
		<	<					158.38904
surface	oppervlakte	NN>	N>	2.00	2.099.709	0.03244	24	
		<	<					93.14844
volcanoes	vulkanen	NNS>	N>	3.00	1.929.182	0.01009	8	
		<	<					114.58168
fault	breuklijn	NN>	N>	1.00	1.761.849	0.01550	12	
		<	<					114.58168
fault	verschuiving	NN>	N>	1.00	1.761.849	0.01550	12	
		<	<					114.58168
fault	breuk	NN>	N>	1.00	1.761.849	0.01550	12	
		<	<					125.79682
miles	kilometer	NNS>	N>	2.00	1.760.330	0.03317	24	
		<	<					125.79682
miles	meter	NNS>	N>	2.00	1.760.330	0.03317	24	
		<	<					92.62205
plates	platen	NNS>	N>	2.00	1.680.587	0.01346	12	
		<	<					73.90449
crater	krater	NN>	N>	2.00	1.660.742	0.00865	7	
		<	<					73.90449
crater	kraterwand	NN>	N>	2.00	1.660.742	0.00865	7	
smoke-shrouded		<	<					0.00000
rim		JJ NN>	ADJ N>	2.00	1.609.905	100.072	1	
		<	<					99.88421
heat	hitte	NN>	N>	1.00	1.589.817	0.02019	15	
		<	<					99.88421
heat	warmte	NN>	N>	1.00	1.589.817	0.02019	15	
		<						0.00000
fourty-five miles		JJ NNS>	< ADJ N>	2.45	1.581.983	100.072	1	

		<						0.00000
five-and-half miles		JJ NNS>	< ADJ N>	2.83	1.581.983	100.072	1	
		<						0.00000
storm-tossed seas	woelig water	JJ NNS>	< ADJ N>	1.73	1.570.367	100.072	1	
never-ending movement		< JJ NN>	< ADJ N>	3.46	1.544.743	100.072	1	0.00000
		<						0.00000
tectonic plates	tektonische platen	JJ NNS>	< ADJ N>	2.45	1.522.137	500.361	6	
		<						0.00000
tectonic plates	platen	JJ NNS>	< ADJ N>	2.45	1.522.137	500.361	6	
All-year-round swimming		< JJ NN>	< ADJ N>	2.00	1.502.432	100.072	1	0.00000
		<	<					57.19543
molten		NN>	N>	2.00	1.502.266	0.00469	5	
		<	<					69.23534
stadium	stadion	NN>	N>	2.00	1.368.531	0.01009	8	
		<	<					57.41363
earthquake	aardbeving	NN>	N>	2.00	1.347.970	0.00865	7	
		<	<					57.41363
earthquake	aardbevingen	NN>	N>	2.00	1.347.970	0.00865	7	
		<	<					91.59002
aquifer	aquifer	NN>	N>	3.00	1.344.635	0.00433	4	
		<	<					59.62294
earthquakes	aardbevingen	NNS>	N>	3.00	1.334.042	0.00577	5	
		<	<					45.09058
hangar		NN>	N>	2.00	1.327.002	0.00577	4	
		<	<					52.13947
cracks	scheuren	NNS>	N>	1.00	1.308.067	0.00721	6	
		<	<					52.13947
cracks	scheurtjes	NNS>	N>	1.00	1.308.067	0.00721	6	
		<	<					72.74440
rock	gesteente	NN>	N>	1.00	1.293.543	0.02019	16	
		<	<					72.74440
rock	rotsen	NN>	N>	1.00	1.293.543	0.02019	16	
		<	<					46.88440
turtles	schildpadden	NNS>	N>	2.00	1.271.215	0.00577	5	
		<	<					46.88440
turtles	tank	NNS>	N>	2.00	1.271.215	0.00577	5	
super-heated molten rock		< JJ NN>	< ADJ N>	2.00	1.266.620	158.544	1	0.00000
								0.00000
magnetic field	magnetisch veld	< JJ NN>	< ADJ N>	1.73	1.138.278	900.649	10	

								0.00000
magnetic field	veld	< JJ NN>	< ADJ N>	1.73	1.138.278	900.649	10	
								0.00000
virtual earth	virtuele aarde	< JJ NN>	< ADJ N>	1.41	1.138.108	800.577	8	
		<	<					27.55630
celcius		NN>	N>	2.00	1.115.632	0.00288	2	
		<	<					43.32602
magma	magma	NN>	N>	2.00	1.110.958	0.00288	3	
		<	<					39.69078
turtle	schildpad	NN>	N>	1.00	1.109.555	0.00577	5	
		<	<					48.40305
core	kern	NN>	N>	1.00	1.070.848	0.00961	10	
		<	<					34.30087
volcano	krater	NN>	N>	3.00	1.056.506	0.00433	4	
separate								0.00000
free-floating blocks		< JJ JJ NNS>	< ADJ ADJ N>	2.08	1.050.332	158.544	1	

4.2 The student experiment

4.2.1 Experimental set-up

Twelve Master's students with average to good marks, enrolled in an English-Dutch translation course, agreed to take part in the proof-of-concept translation experiment.

The English source text the students had to translate into Dutch – with and without glossaries as will be explained below - was a selection of sentences from the voice-over commentary scripts of the experimental corpus discussed in Chapter 2 (*The Earth Machine – Land, Madagascar – The Island of Marvels* and *The Secret World of Pain*'), totalling 775 words. Each sentence contained one or more different terms. 'Term' means figuring in the glossaries because 'manually labelled as such in the GS or extracted by the automatic system TExSIS (TX)'. Table 17 below provides the number of term types (unique terms) in the GS only, in TX only and, in the last column, the term types extracted by both glossaries. The column 'only in TX' indicates that merely one term was extracted by the TX only and not by the GS. In other words, the GS glossary contains more terms than the TX glossary. Moreover, the total number of words in the first column illustrates the 'term density' in these texts.

Table 17 Number of term types per text in the glossaries

No of term types → Episode (total No of words) ↓	only in GS	only in TX	GS and TX
Earth Machine (234)	16	1	23
Madagascar (222)	31	0	17
Pain (319)	34	0	16

Table 18 shows the number of term tokens (total number of terms, repetitions included) per text in the GS and in the TX glossary.

Table 18 Number of term tokens per text in the glossaries

No of term tokens → Episode (total No of words) ↓	GS	TX
Earth Machine (234)	39	24
Madagascar (222)	48	17
Pain (319)	50	16

For a correct comprehension of the text, additional information was added where needed in italics (e.g., the adjunct to which a relative pronoun refers or the preceding sentence). All the students translated the same source text from English into Dutch, presented on their desktop in a Microsoft Word document. Appendix 3 includes the source texts of the experiment.

The participants did not make use of the video as the source text consisted of self-contained clips with no coherence between them. The terminology was clear from the verbal context and there was no need to match timing or style with the images and intonation of the voice talent for the purpose of this study. Furthermore, even in daily practice, audiovisual translators do not always have access to the video, for example, in rush jobs and/or for copyright reasons. A last reason for not providing the images is merely practical. It would take too much time to have the participants watch three documentaries. Cutting the footage that corresponds to the source text sentences and edit it into one video is only feasible if one has the copyrights to the film. Moreover, the retrospective survey of the professionals' experiment (discussed in Section 4.3) revealed that only one of the six audiovisual translators participating in the experiments, declared that for one of the tasks the video could have been useful, two participants said the video might have speeded up the translation process and three participants said working without the images made no difference for this task.

In a first session, all the participants translated without bilingual glossary but they were allowed to consult all the digital sources available to them, including dictionaries from the university website.

In a second session, which was held two months later (the span I expected the participants would need 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 done for the first session had been assessed, so that two comparable groups could be made. The first group was asked to translate the same text with the GS glossary, while the second group did the translation using the TX glossary. The glossaries were presented on their desktop in a Microsoft Excel document, in alphabetic order, by analogy with dictionaries. While the students were translating, Inputlog[10] (see Section 3.3) registered the whole translation process. 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 due to technical failure would be limited to maximum one text.

It so happened that, due to human and 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. In a trial run involving one student, the technical and organisational efficacy of this set-up was tested. On the basis of this trial run, the time needed for translating the source text was estimated to be two and a half hours. Appendix 4 provides the intructions the students received before starting the experiment.

4.2.2 Data analysis and results

4.2.2.1 Process time and pause time

First, the question as to whether or not the integration of a domain-specific, bilingual glossary in the translation process reduced the translators' process time and pause times before terms was addressed.

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 (see Table 19). In the first session, when working without glossaries, only half of the students were able to finish the three texts. Six students translated only 31-64% of the last text. For that reason,

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

the last text was not included in the statistical analyses. Yet, in the second session, when the glossaries were used, all the students finished the task within the time given.

Table 19 Average process time rounded off to the nearest second for both groups in both conditions

Average process time in seconds	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[11] for using this test were satisfied. On average, the participants spent significantly more time on the two texts when working without glossary than they did with glossary. The average process time difference for the TX group for both texts was 524.29 s (C.I. of the difference 259.30 to 789.29 s, p = .001) and for the GS group 920.92 s (C.I. of the difference for 496.22 to 1345.63 s, p = .001).

Next, the difference between the pause time in the groups, working with and without glossaries, was addressed. The pause time was calculated by means of the Inputlog general analysis, as explained in Section 3.3. The results are illustrated in Table 20 below.

Table 20 Average pause time for terms rounded off to the nearest second for both groups in both conditions

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

Again, the data were analysed in SPSS version 20 with a two-samples-paired t test, 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 TX group was 267.17 s (C.I. of the difference: 154.54 to 379.81, p = .000) and for the GS group 337.79 s (C.I. of the difference 156.45 to 519.13, p = .002). Comparing the average pause time between both conditions in Table 20, we can observe a pause time gain for terminology of 38% for the TX group and 55% for the GS group. A comparison between the average process time

[11] Following assumptions for a two-samples-paired t test need to be satisfied: independent differences between the elements of the sample (no correlation between them) and normally distributed differences. The variances in the two groups (A and B) were not significantly different from each other.

in both conditions (Table 19) reveals a process time gain of 17% for the TX group and 31% for the GS group.

4.2.2.2 Use of glossaries vs use of other resources

In addition to the assessment of the process and pause time in both conditions, the degree to which the students used the bilingual glossaries, the internet and dictionaries (research question 4) was investigated. The hypothesis is that the more exhaustive the glossary is, the less time the translators will spend browsing the internet and dictionaries, because they will rely more on the glossary for the translation of terminology.

The total pause time before terms was divided into two categories: pause time for bilingual glossaries (either GS or TX) 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 3) shows how the two categories were divided over the total pause time before terms in milliseconds. In the case of the GS group, the time spent consulting the glossary and the time spent on dictionaries plus the internet is equally distributed, whereas the TX group spent clearly more time consulting dictionaries or the internet. This confirms the hypothesis above.

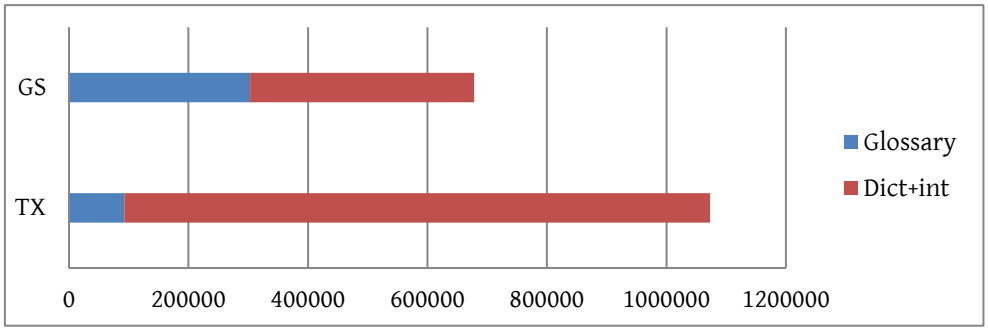

Figure 3 Use of the glossaries vs use of dictionaries and internet per group in milliseconds

4.2.2.3 Terminological errors

Besides the process/pause time and the use of resources, the number of terminological errors was examined: did they decrease or not when working with a bilingual glossary? A terminological error means that a term from the GS was translated with a term not corresponding to a correct translation in this context.

The graph in Figure 4 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 due lack of time, was not considered).

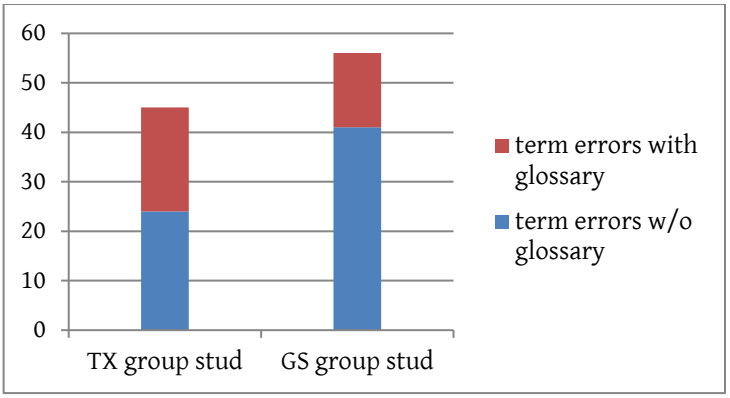

Figure 4 Total number of term errors made when translating with and without a glossary

The correspondingly paired t test showed no significant difference in the number of errors between the two conditions (p= .06), based on data of both the GS group and the TX group, and based on the TX group only. However, a significant difference can be observed when analysing the number of errors of the GS group only, where students translated with and without glossary. On average, participants made significantly fewer terminological errors, working with the GS than without. The average difference was 2.16 (C.I. of the difference: 0.45 to 3.87, p = .01).

4.2.2.4 Retrospective survey

In a final stage, a short retrospective survey was carried out among the candidates, inquiring into their experience of translating with a glossary and the set-up of the experiment. Three questions were put to them:

1) Do you think you worked more quickly with the glossary?
2) Do you think the glossary was beneficial for your translation?
3) In the second session, did you remember parts of the text/terms from the first session?

To question 1, nine said they did, three (from the TX group) said they did not, whereas Table 18 illustrates they all worked more quickly. Question 2 was confirmed by the GS group and denied by the TX group. As for question 3, one student remembered many terms, ten of them remembered some and one did not remember any of them.

As a result of this last finding, the professionals' experiment was improved by extending the time between the two sessions from two months to four months.

4.3 The professional translators' experiment

4.3.1 Experimental set-up

Twelve professional English to Dutch translators, with an average of twenty years of experience (three to thirty-eight), participated in the experiment. Six were audiovisual translators, while the other six had expertise in different kinds of texts, such as financial, scientific, commercial, legal and technical. They were paid 100€ for the two sessions. The experiment was conducted in their own working place with a laptop with Inputlog[12] installed. Appendix 4 provides the intructions the professionals received before starting the experiment.

The participants translated the same English source texts as the students plus one extra text: an excerpt from the documentary *The Hunt of the Higgs* (2012), all presented on their desktop in a Microsoft Word document. This extra text was added in order to raise the number of observations and as a consequence, the quality of the statistical analysis. The total number of words to be translated was 1017 (as opposed to 775 for the students). The participants translated the same texts once with and once without bilingual glossaries. The experiment design is discussed into greater detail below. The glossaries were presented on their desktop in a Microsoft Excel document, in alphabetic order, by analogy with dictionaries.

Table 21 provides the number of term types (unique terms) in the GS only, in TX only and, in the last column, the term types extracted by both glossaries. The column 'only in TX' indicates that only seven terms were extracted by the TX only and not by the GS. In other words, the GS glossary contains more terms than the TX glossary. Moreover, the total number of words in the first column illustrates the 'term density' in these texts. Table 22 illustrates the number of term tokens (total number of terms) per text.

Table 21 Number of term types per text in the glossaries

No of term types → Episode (total No of words) ↓	only in GS	only in TX	GS and TX
Earth Machine (234)	16	1	23
Madagascar (222)	31	0	17
Pain (319)	34	0	16
Higgs (242)	25	6	13

[12] http://www.inputlog.net/ (accessed 20 October 2014).

Table 22 Number of term tokens per text in the glossaries

No of term tokens → Episode (total No of words) ↓	GS	TX
Earth Machine (234)	39	24
Madagascar (222)	48	17
Pain (319)	50	16
Higgs (242)	38	19

As in the student experiment, additional information was given where needed in italics for a correct comprehension of the text and all the professionals translated the same source text from English into Dutch. Appendix 3 includes the fourth source text *The Hunt for the Higgs*. The participants did not make use of the video for the same reasons as the students (self-containing clips, clear terminology, no matching needed with timing or style and voice talent, copyright and time issues). Unlike the student experiment, there was no set time limit for this experiment, so that all the candidates were able to complete the assignment. In this way, all translations could be used for statistical purposes and a maximum number of observations was guaranteed. Due to human failure and, in one case, due to a lack of time, three translations were not registered at all.

The candidates were divided into two equal groups, three audiovisual translators in one group and three non-audiovisual translators in the other.

The study design for this experiment with professionals was improved in three ways:
- no time limit was set to finish the translation task
- the time between the two translation tasks was doubled (four months instead of two), see also 4.1.2.4.
- counterbalancing[13] was introduced.

The study design was set up as follows. The participants translated the same text twice, once with the glossary and once without. In order to control the memory effect, despite the four months' interval between the two sessions, counterbalancing was introduced. Half of the participants (group A) were asked to work without the glossary and the other half with in the first session. For the second session, this working condition was inverted: group A worked with the glossary and group B without. Hence, both groups had the advantage of the memory effect in the second session, however, group A also had the advantage of the glossary in this second session and not in the first one whereas group B had the advantage of the glossary in the first session and not in the second one. This

[13] Counterbalancing can be defined as using all of the possible orders of conditions to control order effects. (Cozby, 2009)

means that the advantage of the memory effect was equally distributed over both conditions: working with and without the glossary.

In terms of mathematics, counterbalancing can be explained as follows:

(A) In the first condition, working with the glossaries in session t^2 and without in t^1, the process time for translator$_i$ =

PROCESS TIME FOR t^1	PROCESS TIME FOR t^2
$(x_i + e_i^1)$	$(x_i + e_i^2 - M_i - G_i)$

x_i = average process time a translator needs, without memory effect, without glossaries

e_i^1 = a random value added for t^1 because the process time for a translator will never be exactly the same

M_i = the memory effect providing a time benefit for translator$_i$ on t^2

G_i = the benefit (can also be negative or zero) translator$_i$ obtains with the glossaries

The difference between both process times = $e_i^1-e_i^2+M_i+G_i$

(B) In the second condition, working with the glossaries in session t^1 and without in t^2, the process time for translator$_i$ =

PROCESS TIME FOR t^1	PROCESS TIME FOR t^2
$(x_i + e_i^1 - G_i)$	$(x_i + e_i^2 - M_i)$

The difference between both process times = $e_i^1-e_i^2+M_i-G_i$

Considering $d_i = e_i^1-e_i^2$, the random difference value between session t^1 and t^2:

Working with the glossaries in session t^2: the difference value = $d_i+M_i+G_i$

Working with the glossaries in session t^1: the difference value = $d_i+M_i-G_i$

In other words: in the difference variables, the 'between subject' variables (= x_i) and their corresponding variances have been cancelled. Only d_i, the 'within subject' variables are left over. It is common knowledge in statistics that the 'within subject' variances are much smaller than the 'between subject' variances. The memory effect is similar in both groups, although there can be a difference from one translator to another, therefore an index i has been added. Calculating the average process time of all participants, it is clear that the main difference in the statistical difference variables will be because of factor G_i, positive in one group and negative in the other one. The average difference of the statistical difference variables between the two groups is expected to be in the order of twice G_i. Indeed, both groups have the same memory effect whereas the glossary effect works in the opposite way: in the first group it increases the process time differences, in the second group it reduces them.

The retrospective survey (see Section 4.3.2.4 for details) revealed that eleven participants remembered at best the general content and at most one term.

Having the same participants translate twice the same text in a counterbalanced setting enhances substantially the statistical reliability. Only variances 'within subject' have to be calculated because people act as their own control, which means that a smaller sample size is possible. A power analysis of the sample size which is needed to detect a determined effect size is derived from Cohen's effect size indexes (1988).

For the thesis under study, only a large effect size (Cohen's d = .80 as in Cohen 1988:35), i.e. a considerable difference between working with and without the glossaries, is interesting. In each group, there were 24 observations from 6 translators working on 4 texts. A power analysis for this effect size, n = 24 and significance level α = .05 shows that the power = .77 (Cohen 1988: 36). This means that we had a probability of 77 % to detect large effect sizes, which is considered important.

4.3.2 Data analysis and results

4.3.2.1 Process time and pause time

In order to implement this design statistically, all analyses were elaborated with the difference variables: the process time (and pause time) of the first translation minus the process time (and pause time) of the second translation, thereby creating process time differences and pause time differences. Subsequently, to examine whether or not translating with or without a glossary resulted in a significant difference, a comparison was made between group A, those working without a glossary in the first session and with a glossary in the second session, and group B, those working first with and then without a glossary. If working with a glossary was more efficient, I expected group A to have longer process time differences (and pause time differences) compared to group B. In order to analyse whether or not group A has longer process time differences (and pause time differences) than group B, an independent sample t test was used. P-values below 0.05 were considered statistically significant.

At first and as for the student experiment, the question as to whether or not the integration of a domain-specific, bilingual glossary in the translation process reduced professional translators' process and pause times before terms was addressed.

It transpired that the average process time differences and pause time differences were reduced significantly when they worked with the glossary (see Table 23). Group A, working without a glossary in the first session and with a glossary (GS or TX) in the second

session, had longer process (or pause) time differences[14] than group B, working with a glossary in the first session, and without in the second session. The average reduction of the process time differences for both groups was 399.39 s (Confidence Interval of the difference 80.97 to 717.82 s, p. = .015) and the average reduction of the pause time differences for both groups was 113.52 s (C.I. of the difference 30.73 to 196.31 s, p. = .008).

Table 23 Average process and pause time differences in milliseconds for group A and B

	average process time diff.	average pause time diff.
group A (first w/o, then with)	672877	142351
group B (first with, then w/o)	-207484	-25333

Next, the process time differences and pause time differences of the GS and the TX group separately were analysed. On average, for the TX group, both process and pause time differences were significantly reduced when working first with a glossary compared to working first without. The average reduction of the process time differences was 469.75 s (C.I. of the difference 92.86 to 846.64 s, p. = .017) and the average reduction of the pause time differences was 124.75 s (C.I. of the difference 39.47 to 210.04 s, p. = .006). For the GS group, neither pause time differences nor process time differences were significantly reduced. The average reduction of the process time differences was 314.05 s (C.I. of difference -249.17 to 877.26 s, p. = .258) and the average reduction of the pause time differences was 102.61 s (C.I. of the difference -53.41 to 258.64 s, p. = .185).

4.3.2.2 Use of glossaries vs use of other resources

The next aspect to be analysed was the degree to which professional translators use a bilingual glossary and if they consult it more as they get used to it. The total pause time before terms was divided manually into two categories: pause time for bilingual glossaries (GS or TX) 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). Figure 5 shows the pause time before terms per source in milliseconds. In the case of the GS group, the time for consulting the glossary was more than the pause time for consulting dictionaries or the internet, whereas the TX group spent less than one fifth of the pause time for consulting dictionaries or the internet on consult the glossary. In other words, the longer the glossary was, the more time the translators spent

[14] Differences of the process/pause time between the first and the second session.

consulting it, the longer the total pause time and consequently, the longer the total process time.

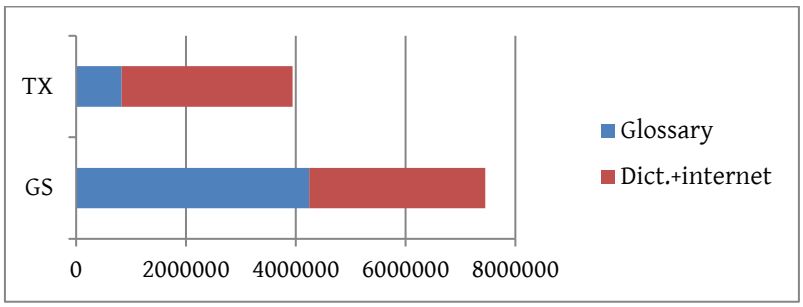

Figure 5 The use of the glossaries vs. the use of dictionaries and internet per group in milliseconds

Furthermore, by means of a line graph (see Figure 6) based on the pause time before terms per text, the evolution was examined towards the end of the assignment in the use of the glossary, the internet and dictionaries. However, no clear answer to research question 5 can be provided. That question seeks to determine to what degree Master's students in translation and professional translators use the glossary and if they consult it more as they become more used to having it at hand. Despite a visual difference between the pause times and between the groups, no conclusion can be drawn since the assignment was done with different texts. Yet, the retrospective survey revealed that the TX group preferred to consult the internet and dictionaries. Three of the six candidates stated that the glossary was not useful and/or they did not trust it.

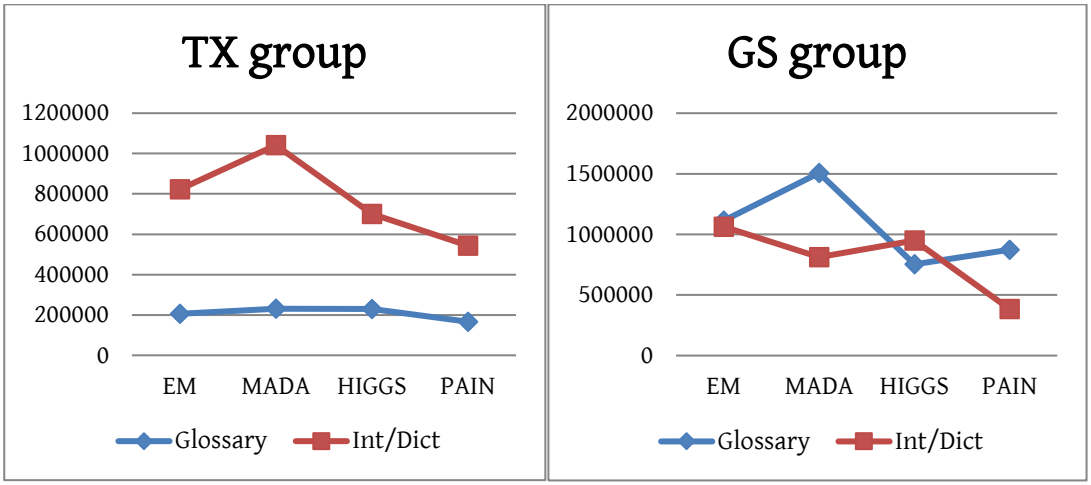

Figure 6 Evolution in the use of the glossary vs. the use of the internet and dictionaries through pause time before terms

In addition, the degree to which the candidates used the glossary was investigated, counting manually how many terms were double-checked (both in the glossary and in a dictionary or the internet), how many terms were looked up only in dictionaries or on the internet (despite the fact that they were listed in the glossary) and how many terms were checked only in the glossary. The average percentage was calculated for each

category. Regarding the search behaviour of the candidates, Table 24 shows that the GS group checks three quarters of the terms only in the glossary whereas the TX group checks one quarter of the terms only in the glossary. The figures in the first column indicate the average percentage of terms double-checked by the candidates (in both the glossary and the internet and/or dictionaries) compared to the total average number of terms they looked up. The second column shows the average percentage of terms the candidates checked only on the internet or in dictionaries, despite the fact they were listed in the glossary. The third column presents the average percentage of terms only checked in the glossary.

Table 24 Search behaviour per group, expressed in average percentages vs. total average number of terms looked up

	Average % terms double-checked	Average % terms only checked in int./dict.	Average % terms only checked in the glossary
GS group	13	10	77
TX group	3	69	28

In order to appreciate the level of difficulty of the terms, Table 25 provides the average % of terms looked up (in the glossary and/or in dictionaries or on the internet).

Table 25 Average % of terms looked up vs. the total average number of terms

	Average % terms looked up vs. total average number of terms
GS group	48
TX group	38

4.3.2.3 Terminological errors

In addition to the process/pause time and the use of resources, whether or not the number of terminological errors decreased when working with a bilingual glossary, was examined. A terminological error means that a term from the GS was translated with a term not corresponding to a correct translation in this context.

A paired-sample t test was used to analyse the statistical difference in term errors between the two working conditions of the GS group and the TX group. A bar graph (see Figure 7) presents the number of errors made by both groups translating with and without glossary.

The analysis showed a significant difference between the number of term errors for the GS group when working with, or without, the glossary. The average difference was 2.67 (C.I. of the difference 0.50 to 4.83, p. = .025). However, no significant difference was

noted in the TX group. The average difference was 0.00 (C.I. of the difference -2.89 to 2.89, p. = 1.00).

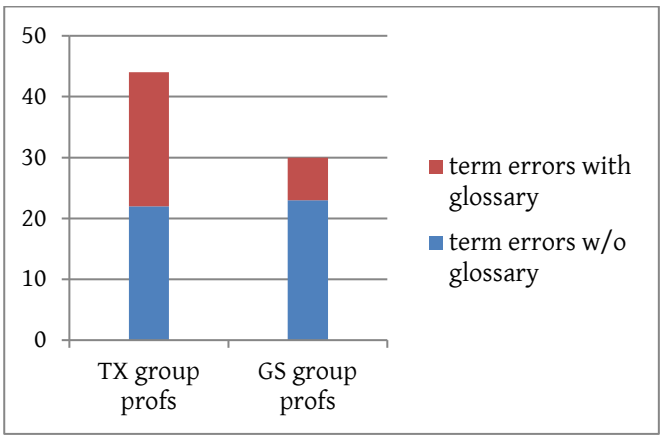

Figure 7 Number of term errors made when translating with and without a glossary

4.3.2.4 Retrospective survey

In a final stage, a short retrospective survey was conducted among the candidates, inquiring into their experience of translating with a glossary and the set-up of the experiment. Four questions were asked:

1) Did you trust the glossary?
2) Do you think the glossary was beneficial for your translation?
3) In the second session, did you remember parts of the text/terms from the first session?
4) Do you think translating without video was a handicap? (the six audiovisual translators, not the others, were asked this question)

The answer to question 1 was compared with the search behaviour for double-checking.

The survey showed that five of the participants trusted the glossary, whereas seven said they tended to double-check the terms on the internet or in dictionaries or even preferred to work without the glossary and search for the correct translation on the internet. However, when asked whether or not the glossary helped them in their translation, nine of the 12 candidates responded it did, mentioning time-saving and verification in case of doubt as their main reasons for using it. Questioned after the second session about what they remembered from the text of the first session, with one exception (see below) all candidates stated they remembered at best, the general content, no terms or at the most only one. One participant said she remembered several terms.

The six audiovisual translators were also asked if they thought it was a handicap to work without video. Two of them replied they might have worked a little bit quicker with the video, but the product would have been the same. One mentioned that for one of the four texts, images would have clarified some of the content but not for the other texts.

The other three candidates stated that for this assignment, it made no difference whether they had the video or not.

4.4 Summary

The present chapter first explained how the bilingual glossaries for the translation experiments were drawn up.

Next, the translation experiments with students and with professional translators were addressed with respect to three issues: the pause and process time, the use of the glossaries and the terminological errors. On the basis of these issues, it was possible to determine the impact of domain-specific, bilingual glossaries on translators' workload and workflow.

1) The pause and process times:

The table below illustrates that for all pause and process times, the statistical results indicate a significant difference between working with and working without a glossary, except for the pause and process time of the professionals' GS group.

Table 26 Results of the statistical analyses for the pause and process times of the experiments

		TX + GS	TX	GS
Pause time	Students	Significant difference	Significant difference	Significant difference
	Profs	Significant difference	Significant difference	No significant difference
Process time	Students	Significant difference	Significant difference	Significant difference
	Profs	Significant difference	Significant difference	No significant difference

2) The use of the glossaries:

For both the students and the professionals, the GS group consulted the glossaries and the internet + dictionaries for about the same amount of time, while the TX group consulted its glossary far less than the other sources.

For the professional translators, the evolution in the use of the sources was analysed. A visual difference can be noticed between the pause times and between the groups. Yet, no conclusion can be drawn as the assignment was done with different texts.

Furthermore, the data analysis reveals that the GS group checked ¾ of the terms in the glossaries only, while the TX group checked ¾ of the terms on the internet or dictionaries only.

3) Terminological errors:

Both students and professional translators working with the GS glossary made significantly fewer errors compared to the students and translators working without the glossary. For the TX groups of students and professionals, there was no significant difference between the two conditions.

4) Retrospective survey:

The overall conclusion of the survey is that the glossaries were considered time-saving and useful for verification in case of doubt by most of the candidates.
All these results will be discussed in Chapter 5.

Chapter 5
DISCUSSION

The main objective in the thesis under study was to establish to what extent a bilingual glossary of domain-specific terminology improves the efficiency of an audiovisual translators' work, translating the off-screen commentary of documentaries. In doing so a number of additional objectives of methodological nature were explored, fine-tuned and evaluated. These will be considered in the present discussion.

Chapter 1 covers the literature review. The first issue addressed in Section 1.1 (p. 5), is what constitutes a documentary film in general and for the Flemish public broadcaster VRT, who made the corpus available, in particular. The brief historical introduction is needed to better understand the concept of documentary film and its ambiguity, viz. a hybrid genre containing both creative and technical aspects.

Documentaries are audiovisual products associated with realism but the discussion about what is to be considered as a documentary and what not is an ongoing issue. Many scholars focus on the fuzzy boundaries between fiction and non-fiction. In the last two decades, however, digitalisation has brought film-makers to expand the traditional boundaries of content, narration techniques and distribution, bringing documentaries to a multifaceted genre, in continuous search for innovation and influenced by different contexts. Therefore, the most interesting approach is that of Chanan (2007): an extended family tree with different branches as clusters of conventions for all styles and techniques.

At the same time, however, this implies that a study into "documentaries" and their translations is somewhat problematic. If what constitutes a documentary can vary greatly, the challenges for translation and the tools to be used will vary to the same degree. In the case of this dissertation, the presence and specificity of terminology is a crucial feature that required additional analyses for the selection of both the corpus and the terminology extraction system (see p. 49).

Not only are definitions very broad, few classifications exist for documentaries and the above-mentioned new trends have overruled the most influential one by Nichols (1991),

based on representation modes (see p. 11). Consequently, a subject-based classification using the VRT corpus was made, distinguishing the following categories: arts, current affairs, history, lifestyle, music, wildlife, science, society, sports and travel. This classification will be useful to select a relevant corpus i.e. containing domain-specific terminology. The hypothesis was that, while all these subjects can contain terminology, the specificity and number of terms – important features for the performance of automatic terminology extraction systems – will depend upon the subject (wildlife or science documentaries will contain more terminology than lifestyle or travel films) and the target audience (documentaries for people in search of information will use a more specialised language than those made for entertainment).

A last feature that determines the selection of the corpus is the narration technique. Documentaries can use different techniques such as spontaneaous or prepared interviews, staged scenes with dialogues and voice-over commentary. For the purpose of the present study, the voice-over commentary i.e. the voice the viewer hears without seeing the speaker, was selected. The translation of such commentaries is called 'off-screen dubbing'.

The next step in the research, described in Section 1.2 (p. 22), tackles the translation process and product of documentaries bearing in mind that terminology is the key issue in this dissertation. Translating documentaries is a challenging task which requires an all-round knowledge of text types and functions, language registers and translation modes. In addition, examining the literature it became clear that a specific, lexical feature of documentaries is terminology, which is a challenge in itself for translators, which is, of course, the main reason for this research into the effectiveness of domain-specific, bilingual glossaries for reducing the workload of audiovisual translators.

Section 1.3 (p. 34) deals with the underlying technology of terminology extraction systems in order to compare existing systems and select the best one for the thesis under study. Terminology extraction systems can have a linguistic (based on term formation patterns), a statistical (based on quantifiable characteristics of term usage) or a hybrid approach (combining both linguistic and statistical approaches) to detect terms. They can be carried out in a monolingual or in a bilingual setting. Bilingual systems first align the sentences of source and target text. Next, word alignment - the process of statistically matching up words within pairs of aligned sentences - is done. For the extraction of the bilingual glossaries, needed for the present research, systems with different approaches were tested and discussed in Chapter 3.

Chapter 2 presents the corpus of all documentaries broadcast by VRT between 2005 and 2013, and made available by the Flemish public broadcaster. It consists of the English source texts of the voice-over commentary and their Dutch translations, the off-screen dubbing. The analysis of the translations demonstrated that the target texts are free translations as required by the VRT guidelines and determined by the audiovisual mode (see also Chapter 2.4). Yet, these free translations are a drawback for automatic systems

that are based onto term formation patterns and statistical filters like frequency, to distinguish terms from non-terms and/or linguistically motivated aligned chuncks in the case of a bilingual approach. The discussion of Chapter 3 below elaborates on this issue.

The documentaries were classified per subject, upon an ad-hoc classification: arts, current affairs, history, lifestyle, music, wildlife, science, society, sports and travel. The science documentaries were subdivided into two more categories viz. 'human body' and 'earth and space'. A preparatory corpus of ten 'wildlife', 'human body' and 'earth and space' documentaries for different target audiences was selected in order to investigate whether or not this text type contains domain-specific terms. A manually labelling of the terms revealed that the texts for an audience in search of information contained more domain-specific terms than those for an audience in search of entertainment, confirming so the hypothesis mentioned on page 17.

Out of these ten documentaries, an experiment corpus of three texts was selected (one for each subject) giving priority to the specificity and not to the frequency of the terms. With this experiment corpus the core of the present dissertation i.e. to understand whether or not the integration of automatically extracted bilingual glossaries in the translation process improves the efficiency of an audiovisual translators' work, was examined.

Chapter 3 fine-tunes the methodology used to reach this purpose.

The first step was the creation of a gold standard as an objective means for testing three automatic term extraction systems with different underlying technology (see Section 1.3.4.3): SDL Multiterm Extract 2011 Trados®, Similis® and TExSIS[1]. Three annotators labelled the domain-specific terminology from the experiment corpus manually. By means of precision, recall and F-score, the inter-annotator agreement was calculated. The F-scores varied from 35.74% to 63.13%, suggesting a strong disagreement between the annotators which demonstrates the fuzzy boundaries between what is to be considered a 'term' and a 'non-term'. For translational purposes of the present - hybrid - genre, a broad interpretation of what a term is, is more useful so the GS was created assembling the terms of all three annotators into one glossary without any filtering.

The second step consisted of testing the three term extraction systems. SDL Multiterm Extract 2011 Trados®, Similis® and TExSIS extracted the terminology from the experiment corpus and precision, recall and F-score were calculated. TExSIS proved to be the best performing system. Yet, even though its precision scores approached those of term extraction tests of technical texts (see Section 3.2.3), its recall scores came short of the expectations, which means that the system failed to extract many terms compared to the GS. The reason for this is twofold. Automatic systems need enough correspondence

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

between script and translation to be able to associate terms of the source text with their translation whereas the target texts in the corpus under study were free translations (see Chapter 3). In addition, as explained in Section 1.3.4, frequency undoubtedly characterises terms. Since the glossaries were extracted out of only three documentaries, the statistical filters and word-by-word alignment – based on frequency – failed to extract several domain-specific terms. For the glossaries used for the experiments, no filtering was applied to distinguish terms from non-terms as slightly fuzzier glossaries can be useful for the translation of the technical and creative elements in documentaries.

The reason only three documentaries were used is merely practical. Although source and target text of all 181 documentaries made available, have been aligned, it proved too labour-intensive to make a GS out of the aligned texts. In the conclusion, I point out suggestions for further research regarding this issue.

The third step included the translation experiments: a pilot study with Master's students in translation and an experiment with professional translators. Details about the experimental set-up will be discussed in the next section. Logging the translation process and product through Inputlog[2], a keystroke logging software developed at the University of Antwerp, the support of bilingual glossaries was examined. Specific features of the software allowed to determine the participants' total translation time (called hereafter 'process time') and their 'pause time before terms'[3], both indicators of translation efficiency and speed. The statistical analysis of these data aimed at investigating whether or not the integration of a domain-specific, automatically and manually generated bilingual glossary into the translation process reduces the process and pause time before terms. Besides, the number of terminological errors was calculated in order to understand whether or not the candidates made less errors when working with a glossary. In addition, the pause time before terms was used to investigate to what degree the candidates consult the glossaries and to understand if they consult it more becoming used to it towards the end of the task. A retrospective survey added interesting context information to the output of the experiments, allowing a more accurate interpretation of the results.

Chapter 4 deals with the experiments, starting with the experimental set-up. The research plan comprised two translation experiments, with and without bilingual glossaries, carried out by Master's students in translation and later by professional translators. They were asked to translate part of the experimental corpus with and without glossaries, in two sessions, while Inputlog registered their entire translation process. They were divided into two groups: one working with the GS glossaries (the GS group) and another one working with the TExSIS glossaries (the TX group).

[2] http://www.inputlog.net/ (accessed 20 October 2014).

[3] This is the time the participant spent to look up the information needed to translate a term.

In order to obtain reliable statistical results from of the experimental data, as many candidates as possible were needed. Thanks to a long targeted lobbying twelve students and twelve professionals, carefully chosen in terms of skills and experience, were found willing to participate. This may not seem a sample size large enough to detect statistically valid effects, however, a power analysis carried out on the professional translators' data showed an important probability to detect large effect sizes (see 4.2.2.1). Moreover, both parametrical and non-parametrical tests on the pause and process times before terms were carried out. The results of the parametrical tests corresponded with the results of the non-parametrical tests, meaning that the small sample size has not affected the statistical results.

The results of the experiments showed similar trends between the students and the professionals despite modifications introduced in the experimental design of the professionals, which will be discussed in the next section. I will first discuss the results.

For both students and professionals, calculating the data of the GS and the TX group together, all pause and process times reveal a significant difference between the condition in which they worked without the glossary and the one in which they worked with a glossary. This means that a domain-specific bilingual glossary in general (manual or automatic) reduces the working time of translators. Analysing the data of the GS and the TX groups separately students' and professionals' results showed again a significant difference between the two conditions. An exception was the GS group of the professionals where no significant difference was registered between the pause and process times working with or without glossaries.

Considering that the professionals' data are more accurate (I refer to the section below for the improvements in the professionals' experiment design) it appears that the longer the glossary is, the more time translators spend consulting it. As explained in Chapter 3 above, the GS glossaries contained far more term types, more than double, than the TX glossaries. The hypothesis is that if audiovisual translators get used to working with a glossary, their efficiency in consulting it might increase. The evolution in the use of the glossaries by the professionals was calculated just to have an idea of the effect of experiencing the glossaries, however, no conclusion could be drawn this time as the assignment was done with different texts (but see below in the conclusion under further research).

Next, the use of glossaries versus other online sources was investigated. When comparing the pause time before terms of the search in the glossaries with the pause time before terms of the search in the internet or in dictionaries, it turned out that GS group of both students and professionals consulted the glossaries as much as other sources while the TX groups consulted its glossaries far less than other sources. The retrospective survey explained that some of the professionals stopped consulting the TX glossaries as they noted important gaps. Probably for the same reason, the professionals working with

the automatic glossaries checked ¾ of the terms only in the internet or dictionaries while those working with the GS checked ¾ of the terms only in the glossaries. Worth while mentioning is that most translators declared in the retrospective survey that the glossaries were time-saving and useful for verification in case of doubt.

A last issue concerns terminological errors. In order to understand whether or not the integration of glossaries in the translation process improves the efficiency of an audiovisual translators' work, the terminological errors made when working with and without glossaries were counted. The results reveal that both GS groups made significantly less terminological errors while there was no significant difference for the TX groups. As the GS glossaries were more exhaustive than the TX glossaries, they provided a better support in terms of accuracy.

As mentioned before, the students' experiment was designed as a pilot study. It turned out that the experimental design showed room for improvement in relation to three aspects which were immediately implemented in the subsequent experiments with professionals.

Firstly, a time span of two months between the two experiments (first with and then without glossaries) seemed to be too short as one of the candidates remembered most of the terms in the second session. For the professional translators, a time span of four months was taken. When asked in the retrospective survey whether or not they remembered the terms, two candidates said they vaguely remembered only one or two terms, the others did not remember any term.

Secondly, the time limit of two and a half hours to finish the task was too short. In the first session (without the glossaries) half of the students did not finish the last text. Therefore, the data of this one text could not be taken into account for statistical analyses. For the professionals, it was decided not to set any time limit so all data could be included for statistics and the largest sample size possible was reached.

Thirdly and most importantly, counterbalancing i.e. an equal distribution of the memory effect over both conditions (translating with and without the glossaries), was introduced into the professionals' experiment. As a consequence of this counterbalancing, the memory-effect was controlled and did not influence the results. The experiments of the students should have been conducted in the same way, with counterbalancing, however, the importance of the memory effect was underestimated. As a consequence, it was corrected in the professionals' experimental design.

As the discussion indicates, the experiments yielded interesting results with respect to the main research question, the research also yielded interesting insights for future work, especially in terms of the methodology to be used and the corpus to be assembled. Recommendations for further research, tackled in the conclusion below, therefore consider possible improvements of the experimental design and suggestions regarding corpora.

Chapter 6
CONCLUSION

The discussion of the corpus under study and that of the terminology extractions systems revealed two important aspects to be considered in future: one is the translation 'style' (intending here the free translation) and the other one is the size of the corpus.

The natural science documentaries of the VRT corpus in the present thesis do contain domain-specific terminology, as illustrated in Chapter 3. The automatic terminology extraction systems however, show room for improvement in recognizing the terms, most probably due to the free translation and the small corpus, as explained in the discussion of the methodology. Changing the translation style into a more literally translation in order to facilitate automatic term extraction is out of question because of the specific requirements imposed for audiovisual translation (see discussion Chapter 2). Yet, it is possible to enlarge the corpus. The 181 aligned documentaries[1] mentioned in the discussion constitute a good starting point for testing whether or not automatic term extraction systems perform better with a larger corpus of off-screen dubbing than the one used in this research. As discussed before, it is too labour-intensive to make a GS out of such a large corpus but comparing precision and recall of different automatic systems between each other will reveal which system performs best. The glossary extracted by this system can then be used for translation experiments aiming at similar goals as those in the thesis under study. An analysis of the pause and process times of the candidates, working with and without glossaries, combined with a retrospective survey will determine the degree to which these glossaries of parallel corpora are time saving and quality enhancing for the translation of off-screen dubbing.

The glossaries can also be improved adding concordances to the terms. In the glossaries used for the experiments, the terms figured in isolation because it was decided to first examine glossaries with only terms. If the sentences of the source and target texts

[1] About one million words source and target texts together.

are aligned (which is the case for the corpus under study), a bilingual concordance tool can retrieve context and add it to the term. Bowker (2011) emphasises that usage information such as context will help translators to produce a target text that reads well. With this enriched glossaries, new translation experiments can be carried out to investigate the differences in pause and process time working with and without glossaries.

In addition to the VRT corpus, other corpora might be useful to obtain glossaries with higher precision and recall figures. Investigating the terminological challenges in English documentaries to be translated into Catalan, Matamala (2009b) suggests to consult in case of translation problems, besides the image and the spoken discourse inside the documentary, all kinds of specialised reference works including parallel corpora.

In case of natural science, an interesting parallel corpus is constituted by the *National Geographic* articles. They cover both wildlife and science, the topics of the source texts for the experiments in this dissertation, are archived as from 1888 and have been translated into more than 35 languages[2]. Moreover, science popularisation[3] as the text type in the present corpus is close to the popular science language register of the *National Geographic.* Another similar source is the World Wildlife Funds'[4] online publications of parallel and comparable texts. A last suggestion regarding corpora is to consider the OPUS project[5], that provides open parallel corpora of translated texts. It is a growing collection of many different topics and corpora e.g. the European Medicines Agency documents (Tiedemann, 2009). Currently, wildlife is not covered in OPUS but the project is worthwhile to keep into consideration for future work.

Interesting research questions for these corpora concern both terminology and glossaries. A first question can be: "Is the terminology used in parallel corpora of popular science journals specific enough to be detected by automatic terminology extraction systems?" Next, an additional research question regarding the translation of documentaries can be formulated: "Does the integration of an automatically generated bilingual glossary out of popular science journals into the translation process of off-screen dubbing reduce the translators' workload?"

Besides testing these corpora through the questions above, the evolution in the use of a glossary (compared to other sources) is an important issue to investigate. As highlighted in the discussion, this aspect should be addressed carrying out a translation experiment with an extract of off-screen dubbing of the same documentary[6]. The domain-specific

[2]http://help.nationalgeographic.com/customer/portal/articles/1446596-is-national-geographic-magazine-available-in-any-languages-other-than-english- (last consulted on 28 December 2016).

[3] See Section 1.2.2.

[4] http://wwf.panda.org/about_our_earth/all_publications/ (last consulted on 28 December 2016).

[5] http://opus.lingfil.uu.se/ (last consulted on 28 December 2016).

[6] About a 1000 words.

glossary to conduct the experiment should be extracted out of one of the above mentioned corpora. The pause times before terms will illustrate whether or not the candidates use the glossary more if they get used to it. The results can also illustrate whether or not the length of the glossary influences the translation process in terms of time and quality and whether or not the translators' efficiency increases consulting it.

In any case, the implemented design used for the professional translators in the thesis under study has proved to be ecologically valid. Consequently, it can be applied in future research that aims at testing automatic domain-specific glossaries for the translation of off-screen dubbing.

The initial impulse of this dissertation was Matamala's research on terminology in documentaries (2009a, 2009b, 2010) and her suggestion that it must be possible to support the audiovisual translators' workload introducing automation into their work process of translating documentaries[7]. The present study has made a first step in this direction, proving that automatic terminology extraction systems support the workload of audiovisual transIators. Important conclusions in terms of experimental design and corpora, as well as suggestions for future research have been formulated.

In the meanwhile, the automation of translation is rapidly evolving. The future research proposed above will have to be conducted in the context of existing systems at the moment of working and in close cooperation with language technologists. Doing so, corpora and systems can be tested thoroughly. Where needed, systems can be enhanced focussing on this particular kind of translations. Eventually, a ready-to-use tool can be developed for audiovisual translators of documentaries.

[7] Personal conversation during the "Media for all Conference5" in London (2011).

Bibliography

Allard, M. G. P. (2012). Managing terminology for translation using translation environment tools: Towards a definition of best practices (Unpublished doctoral dissertation). University of Ottawa, Ottawa, Canada.

Amparo, A. (2008). Translation technologies. Scope, tools and resources. *Target, 20*(1), 79-102. doi 10.1075/target.20.1.05alc

Agost, R. (1999). *Traducción y doblaje: Palabras, voces y imágenes.* Barcelona: Ariel.

Barbash, I., & Taylor, L. (1997). *Cross-cultural filmmaking. A handbook for making documentary and ethnographic films and videos.* Berkeley and Los Angeles (CA): University of California Press.

Barnouw, E. (1993). *Documentary: A history of the non-fiction film.* Oxford: Oxford University Press.

Bourigault, C., & Jacquemin, C. (1999). Term extraction and term clustering: An integrated platform for computer-aided terminology. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics.* Paper presented at EACL '99, Bergen, Norway. 15-22. Stroudsburg, PA, USA: Association for computational linguistics. doi 10.3115/977035.977039

Bowker, L. (2008). Terminology. In M. Baker & G. Saldanha (Eds.). *Encyclopedia of Translation Studies (2nd ed.).* London/ New York: Routledge, 286-290.

Bowker, L. (2011). Off the record and on the fly. In A. Kruger, K. Wallmach & J. Munday *(Eds.). Corpus-based Translation Studies: Research and Applications*. London/New York: Continuum, 211-236.

Bowker, L. (2015). Terminology and translation. In H. J. Kockaert & F. Steurs (Eds.). *Handbook of Terminology. Volume 1.* Amsterdam/Philadelphia: John Benjamins, 304-323.

Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceeding of the 29th annual meeting of the Association for Computational Linguistics.* Paper presented at ACL '91, Berkeley, California, USA, 169-176. Stroudsburg, PA, USA: Association for computational linguistics. Doi 10.3115/981344.981366

Bruzzi, S. (2000). *New documentary: A critical introduction.* London: Routledge.

Cabré Castellvi, M.T. (1999). *Theory, methods and applications*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Cabré Castellvi, M.T. (2003). Theories of terminology. Their Description, prescription and explanation. *Terminology, 9*(2), 163-200.

Cabré Castellvi, M.T., Estopà, R. & Vivaldi, J. (2001). Automatic term detection: a review of current systems. In D. Bourigault, C. Jacquemin & M-C. L'Homme (Eds.). *Recent advances in Computational Terminology* (pp. 53-88). Amsterdam/Philadelphia: John Benjamins.

Carroll, N. (1996). *Theorizing the moving image.* New York: Cambridge University Press.

Cattrysse, P. (1995). *Handboek scenarioschrijven.* Leuven/Apeldoorn: Garant.

Champagne, G. (2004). The economic value of terminology: An exploratory study. Unpublished report submitted to the translation bureau of Canada, Public Works and Government Services Canada, 36 pp.
Chanan, M. (2007). *The politics of documentary.* London: British Film Institute.
Chapman, J. (2009). *Issues in contemporary documentary.* Cambridge: Polity Press.
Chaume, F. (2014). *Audiovisual translation: dubbing.* Manchester, UK: St. Jerome Publishing
Christensen, T. P., & 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.
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
Corner, J. (2002). Performing the real. *Television & New Media, 3*(3), 255-269.
Cozby, P. C. (2009). *Methods in behavioral research*. Columbus, OH: McGraw-Hill Higher Education.
Daille, B. (1996). Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans & Philip Resnik (Eds.), *The balancing art: Combining symbolic and statistical approaches to language,* (pp.29-36). Cambridge MA: MIT Press.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 67-74.
Ellis, J. C. (1968). The young Grierson in America. *Cinema Journal, 8* (1), *12-21.*
Ellis, J. (2012). *Documentary. Witness and self-revelation*. New York: Routledge.
Espasa, E. (2004). Myths about documentary translation. In P. Orero (Ed.). *Topics in Audiovisual Translation* (pp.183-197). Amsterdam/Philadelphia: John Benjamins.
EU-bridge final report (2015). Unpublished report.
Franco, E. (1999). Discourse or coincidence? A case of documentary translation. In J. Vandaele (Ed.) Leuven: Katholieke Universiteit Leuven: *Translation and the (Re)location of Meaning. Selected Papers of the CETRA Research Seminars in Translation Studies 1994-1996*, 287-316.
Franco, E. (2000). Documentary film translation: A specific practice? In A. Chesterman, N. G. San Salvador & Y. Gambier (Eds). *Selected Contributions from the EST Congress Granada 1998: Translation in Context* (pp. 233-242). Amsterdam: John Benjamins.
Franco, E. (2001a). Inevitable exoticism. The translation of culture-specific items in documentaries. In: F. Chaume, & R. Agost (Eds.). *La Traducción en los Medias Audiovisuales* (pp. 177-181). Castelló de la Plana: Publicacions de la universitat Jaume I.
Franco, E. (2001b). Voiced-over television documentaries. Terminological and conceptual issues for their research. *Target 13*(2), 298-304.
Franco, E., Matamala, A., & Orero, P. (2010). *Voice-over translation: An overview*. Bern: Peter Lang.
Frantzi, K. T., Ananiadou, S. & Mima, H.(2000) Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries manuscript 3(2),* 155-130.
Frantzi, K. T., Ananiadou, S., & Tsujii, J. (1997) Term identification using contextual cues. In: *Proc. 2nd Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC'97)*. Held at the Int. Joint Conf. on Artificial Intelligence (IJCAI-97). Nagoya, Japan.
Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. *Proceeding of the 29th annual meeting of the association for computational linguistics,* Berkeley, California, USA, 177-184.
Gambier, Y. (2012). The position of audiovisual translation studies. In *The Routledge Handbook of Translation Studies.* Routledge. Accessed on: 17 Jun 2017 https://www.routledgehandbooks.com/doi/10.4324/9780203102893.ch3
Gile, D. (2011). La recherche traductologique: méthodes ou approche? [Research in Translation Studies: methods or approach?]. In: C. Foz, & R. Fraser, *Cartographie des Métodologies en Traduction / Charting Research Methods in Translation Studies* 24(2), 41-64.
Gommlich, K. (1993). Text Typology and Translation-Oriented Text Analysis. In S. E. Wright & D. Leland Wright Jr. (Eds.). *Scientific and Technical Translation, VI* (pp. 175-184).

Gregory, M. & Carroll, S. (1978). Language and situation: Language varieties and their social contexts. London: Routledge and Kegan Paul.

Heid, U., & Spranger, K . (2003). Extracting terminologically relevant contexts from chunked corpora. *TIA-2003 Actes des cinguièmes rencontres Terminologie et Intelligence Artificielle,* Strasbourg : LIIA – ENSAIS, 112-123.

Heylen, K., & De Hertog, D. (2015). Automatic term extraction. In H. J. Kockaert & F. Steurs (Eds.). *Handbook of Terminology. Volume 1.* (pp. 203-221). Amsterdam/Philadelphia: John Benjamins.

Herwartz, R. (2011). When does terminology really pay? *Tcworld, October.* Retrieved from http://www.tcworld.info/e-magazine/content-strategies/article/when-does-terminology-really-pay/

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

Kageura, K. (2015). Terminology and lexicography. In H. J. Kockaert & F. Steurs (Eds.). *Handbook of Terminology. Volume 1.* (pp. 45-59). Amsterdam/Philadelphia: John Benjamins.

Kaufman, F. (2004). Un exemple pervers de l'uniformisation linguistique dans la traduction d'un documentaire : de l'hébreu des immigrants de "Saint Jean" au français normatif de Arte. In *Meta 49* (1) 148-160. La traduction audiovisuelle (TAV), Y. Gambier (Ed.)

Kaufman, F. (2008). Le sous-titrage des documentaires: défis et enjeux de l'établissement du texte de départ. In: *La Traduction Audiovisuelle. Approche interdisciplinaire du sous-titrage* (pp. 69-83). J-M. Lavaur & A. Serban (Eds.). Belgique: Traducto, De Boeck université.

Kay, M., & Röscheisen, M. (1993). Text-Translation Alignment. *Computational Linguistics 19*(1), 121–142.

Kelly, N., & De Palma, D. A. (2009). The case for terminology management. Why organizing meaning makes good business sense. Lowell: Common sense advisory.

Koehn, Ph. (2009). Statistical machine translation. Cambridge: Cambridge University Press.

Kozloff, S. (1988). Invisible storytellers. Voice-over narration in American fiction film. Berkeley (CA): University of California Press.

Kozloff, S. (2000). *Overhearing film dialogue.* Berkeley (CA): University of California Press.

Krings, H. P. (2001). *Repairing texts: empirical investigations of machine-translation post-editing processes*, 5. Kent: Kent State University Press.

L'Homme, Heid, Sager (2004). Terminology during the past decade (1994 – 2004). *Terminology, Vol.9:2, 151-161.*

Lacruz, I., Denkowski, M., Lavie, A. (2014). Cognitive demand and cognitive effort in post-editing. In *Proceedings of the third Workshop on Post-Editing Technology and Practice*. Vancouver: AMTA, 73-84

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.

Lommel, (2005). LISA Terminology management survey: Terminology management practices and trends. *Localization Industry Standard Association.*

Macken, L. (2010). *Sub-sentential alignment of translational correspondences,* PhD dissertation. Antwerp: University press Antwerp.

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

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

Matamala, A. (2009b). Main challenges in the translation of documentaries. In: Diaz-Cintas, J. (ed.) *In so many words: translating for the screen*. London: Multilingual matters. 112-126

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

Matamala, A., Fernàndez-Torné, A. & Ortiz-Boix, C. (2012) 'Technology and AD: The TECNACC project'. *Languages and the media 2012,* Berlin.
McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies. An advanced resource book.* London and New York: Routledge.
Michelson, A. (ed.) (1984) *Kino-eye: the writings of Dziga Vertov* (transl. Kevin O'Brien), Berkeley and Los Angeles: University of California Press.
Nakagawa, H., & Mori, T. (1998). A simple but powerful automatic term extraction method. In *COLIN-02 on COMUTERTERM 2002: Second international workshop on Computational Terminology,* 14, 1-7. Stroudsburg, PA, USA: Association for Computational Linguistics.
Nichols, B. (1991). *Representing reality: Issues and concepts in documentary.* Bloomington: Indiana University Press.
Nichols, B. (1994). *Blurred boundaries: questions of meaning in contemporary culture.* Bloomington and Indianapolis: Indiana University Press.
Nichols, B. (2001). *Introduction to documentary.* Bloomington: Indiana University Press.
Orero, P. (2004). The pretended easiness of voice-over translation of TV. Jostrans 2. 76-96
Ortiz-Boix, C. (2016) 'Post-Editing Wildlife Documentaries: Challenges and Possible Solutions' Hermeneus 18. 269-313 http://www5.uva.es/hermeneus/wp-content/uploads/arti09_18.pdf
Ortiz-Boix, C. & Matamala, A. (2016). 'Post-Editing Wildlife Documentary Films: a new possible scenario?' Jostrans 26. 187-210 http://www.jostrans.org/issue26/art_ortiz.pdf
Osstyn, K. (2010). *VRT Leidraad voor commentaarvertalen 2010.* Unpublished manuscript.
Pedersen, J. (2011). *Subtitling norms for television. An exploration focussing on extralinguistic cultural references.* Amsterdam/Philadelphia: John Benjamins.
Pavel, S. & Nolet, D. (2001) *Handbook of terminology.* Ottawa: Public Works and Government Services, Canada.
Planas, E. (2005). *Similis. Translation memory software.* Paper presented at ASLIB International Conference on Translating and the Computer. 2005, London.
Plantinga, C. (1997). *Rhetoric and representation in non-fiction film.* Cambridge: University of Cambridge Press.
Ranzato, I. (2012). Gayspeak and gay subjects in audiovisual translation: strategies in Italian dubbing. *Meta 57* (2). 369-384. doi 10.7202/1013951ar
Rayson, P., & Garside, R. (2000). Comparing Corpora Using Frequency Profiling. In *Proceedings ofthe Workshop on Comparing corpora*, 38th annual meeting of the Association *for Computational Linguistics,* 1-6. Hong Kong, China.
Reiss, K. & Vermeer, H. J. (1984). Towards a general theory of translation action. Skopos theory explained. Manchester UK: St. Jerome Publishing.
Remael, A. (1995/6). From the BBC's 'Voices from the Island' to the BRTN's' De President van Robbeneiland. Linguistica Antverpiensia XXIX-XXX, 107-128.
Remael, A. (2003). Mainstream Narrative Film Dialogue and Subtitling. In Y. Gambier (Ed.), *Screen Translation*, special issue of The Translator 9(2), 225 - 247.
Remael, A. (2007). Whose language, whose voice, whose message? Different AVT modes for documentaries on VRT-Canvas television, Flanders. *TradTerm 13*, 31-50.
Renov, M. (1993). *Theorizing documentary.* New York: Routledge.
Ruby, J. (2000). *Picturing Culture. Explorations of film and anthropology.* Chicago and London: The University of Chicago Press.g
Sager, J.C. (1990). *A practical course in terminology processing.* Amsterdam/Philadelphia: John Benjamins.
Sager, J.C. (1998). In search of a foundation: Towards a theory of term. *Terminology,* 5 (1), 41-57.
Spence, L. & Navarro, V. (2011). *Crafting truth. Documentary form and meaning.* New Brunszick, N.J.: Rutgers University Press.
Šukaytitè, R. (2012). Documentaries on the web: a new expanded consciousness, cinematic genre and cultural experience. *Acta Academiae Artium Vilnensis* 67, p129-137.

Taylor, C. (2002). The subtitling of documentary films. *Rassegna Italiana della Linguistica Applicata 34* (1-2). 143-159.
Tiedemann, J. (2001). Can Bilingual Word Alignment Improve Monolingual Phrasal Term Extraction?. *Terminology* 7 (2), 199-215.
Van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths.
Vintar, Š. (2004). Comparative Evaluation of C-value in the Treatment of Nested Terms. Memura 2004 - Methodologies and Evaluation of Multi-word Units in Real-World Applications (LREC 2004), 54-57.
Vintar, Š. (2010). Bilingual Term Recognition Revisited. The Bag-of-Equivalents Term Alignment Approach. *Terminology* 16 (2), 141-158.
Vivaldi, J. & Rodriguez, H. (2007). Evaluation of terms and term extraction systems. *Terminology, 13 (2),* 225-248.
Vu, T., Aw, A.T., Zhang, M. (2008). Term extraction through unithood and termhood unification. In *Proceedings of International Joint Conference on Natural Language Processing,* 631-636. Hyderabad, India.
Warburton, K. (2013). Processing terminology for the translation pipeline. *Terminology*, 19 (1), 93-111.
Ward, P. (2005). *Documentary: The margins of reality.* London and New York: Wallflower.
Werlich, E. (1975). Typologie der Texte. Heidelberg: Quelle & Meyer.
Winston, B. (1995). *Claiming the real: The documentary film revisited.* London: British Film Institute.
Wright, S.E. (1997). Term selection. The initial phase of terminology management. In S.E. Wright and G. Budin (eds) *Handbook of terminology management.* Amsterdam: John Benjamins, 13-24.
Zielinski, D. & Ramirez, Y.S. (2005). Research meets practice: t-survey 2005. An online survey on terminology extraction and terminology management. http://mt-archive.info/Aslib-2005-Zielinski.pdf

Filmography

Appelbaum, J., Nemec, A., Pinker, J. & Rosenberg, S. (Writers), & Katleman, M. (2015-). *Zoo Australia* Coe, P.T. & Niss, J.F. (Producers). USA: Midnight Radio, Tree Line Film, JP Entertainment, CBS Television Studios.

Brown, N. (Director). (2006). Can we save planet earth? Episode 2 [Television series episode]. In J. Bristow (Producer), *Climate Chaos.* UK: BBC/Discovery Channel.

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.

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.

Darke, W. (Producer). *(1996-) Big cat diary.* [Television series episode]. UK: BBC.

Davies, R. (Writer). (2001-2010). *The Private Life of a Masterpiece* [Television series episode]. In Jones, I.M., Winnan, J. & Gold, M. (Producers), UK: BBC.

Flaherty, F. (Writer), & Flaherty, R. (Director). (1922). *Nanook of the North.* R. Flaherty (Producer) & J. Révillon (Executive Producer).

Flaherty, F. (Writer), & Flaherty, R. (Director). (1926). *Moana.* F. Flaherty & R. Flaherty (Producers).

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.

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.

Grierson, J. (Director). (1929). *Drifters.* Empire Marketing Board (Producers). UK.

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.

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.

Hornsby, Gaby (Producer & Director) (2012) 'Horizon - The Hunt for the Higgs', [Television series episode], in Aidan Laverty (Series editor), *Horizon,* UK: BBC2.

Isaacs, M. (Director). (2001). *Lift.* B. Giles (Executive Producer).

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.

Lindholm, T. (Writer), & Koefoed, A. (Director). (2014) *The arms drop.* Denmark, India, UK, Sweden: Fridthjof Film.

Longinotto, K. (2014). *Love is all: 100 Years of Love and Courtship.* M. Atkin, H. Croall, & M. Rosenbaum (Producers). UK & US: Lone Star Productions, Crossover, & BBC.
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.
Nicopoulos, S. (Writer), & Rose, Y. (Director). (2010). *Is the magnetic pole about to flip?* [Motion picture]. Canada: TGA Production/Ideacom International Inc./CNRS Images .
Pandey, M. (Writer), & Kapadia, A. (Director). (2010) *Senna.* UK, France, USA: Universal Pictures, StudioCanal, Working Title Films.
Peire, F. (Writer) & Gooch, N. & Gortmans, T. (Directors). (2012). *Restaurant Inspector.* Carre, I. (Producer). UK: Channel 5.
Summerill, M. (Producer). (2011). Madagascar – Island of marvels [Television series episode]. In M. Gunton (Executive Producer), *Madagascar.* UK: BBC NHU/Animal Planet.
Vrebos, J. (Director). *(1993-). Het leven zoals het is.* Belgium: Kanakna Productions.

Appendix 1

Criteria and guidelines for manual terminology extraction

(See Section 3.2.2.2. p. 59)

1. Termhood:

Each entry in the extracted lexicon should refer to an object or action that is relevant for the domain. "Termhood" is used to express "the degree to which a linguistic unit is related to domain-specific context".

1.1. All words or groups of words can be considered as a term (see also 2.1)
Noun Phrase (NP) E.g. #decelaration effect#
NP + Adj Phrase as an adjunct E.g. #complex projects#
Prepositional Phrase (PP)
Verb Phrase (with or without adjunct) (VP) E.g. #sailing#
#dredging of rock#

1.2. To determine whether or not a unit is domain-specific depends on your own intuition. Therefore, it is useful to add a score from 1 to 3, indicating the degree of 'termhood', where '1' means very domain-specific, while '3' denotes more general terms but still domain-specific.
E.g. "dredging" = #1#; "vessel" = #2#; maintenance" = #3#

1.3. Domain-specific does not necessarily mean 'difficult, specialised or technical language'.
Terms like "gas", "oil", "ship", "tunnel", "bridge", "road" are well-known terms, but are still domain-specific. Domain-specific means that words occur more frequently in that domain than in other domains or in general language.

1.4. Names of places, persons, institutions are indicated by #NE# (= 'named entity')
E.g. #Jan De Nul Group#NE#

1.5. Not to be indicated as a term:
Years, numbers, universal signs (e.g., %, €, $), mathematic formulas

2. Unithood:

Both single-word terms and multi-word terms have to be labelled + single word terms within multi word terms.

2.1. Each term must constitute a syntactical unit (a phrase).

E.g. *Many young people assume responsibility for executing projects and/or following up on vessels whereas their experience is rather limited.*

#executing#

#projects#

#~~executing /projects~~#: is no unit (= VP + NP as separated phrases)

#following up on#

#vessels#

#~~following up /on vessels~~#: is no unit (= VP + NP as separated phrases)

E.g. *Other, more experienced staff must continue to "go the extra mile" to control this process of growth and ensure that our investments, which are the driving force of this growth, yield a good return.*

#staff#

#growth#

#process of growth#: is a unit (NP with NP as an adjunct)

#investments#

#growth#

#driving force#: is a unit (NP with NP as an adjunct)

#yield#

#return#

2.2. As mentioned in 1.2., add a score from 1 to 3, indicating the degree of 'termhood' (also for terms that are part of other terms) : it may occur that complex terms get another score than their 'basis', the single unit.

E.g.: #Raad van Bestuur#Board of Directors#Conseil d'Administration#Vorstands#1#

#Raad#Board#Conseil #Vorstands#2#

2.3. Do not include articles, they are not part of the term.

2.4. For special cases use "***"

- Compounds of more than one term where the common part is used only once, are labelled as they occur in the text (the space in between is indicated by "***")

E.g.: *dredging and land reclamation projects*

#dredging#

#land#

#dredging *** projects#

#land reclamation#

#land reclamation projects#

- Separable verbs (in Dutch):

E.g.: *Bij regenweer vaart het schip niet uit.*

#vaart *** uit*#

3. Tips

- use the copy-paste function to avoid errors
- put everything between ## (before and after every term, between every language, between the score, the NE)
- copy the terms as they occur in the text (declined, conjugated)

E.g.: #Raad van Bestuur#Board of Directors#Conseil d'Administration#Vorstands#2#

Appendix 2

The glossaries

1a. The Earth Machine – TExSIS glossary

aquifer	aquifer
aquifers	aquifers
ball	bol
bay	baai
bed	bodem
bottom	bodem
cabin	hut
centre	middelpunt
centre	kern
centre of the earth	middelpunt van de aarde
chambers	kamers
continental plate	continentale plaat
coast	kustlijn
core	kern
cracks	scheuren
cracks	scheurtjes
crater	krater
crater	kraterwand
crust	aardkorst
crust	korst
degrees	graden
destination	eindbestemming
diamonds	diamanten
drinking	liter
edge	rand
earth	aarde
earth	aardkorst
earth	aardoppervlak
earthquakes	aardbevingen
earthquake	aardbeving

earthquake	aardbevingen
energy	energie
engine	motor
effects	gevolgen
exit	uitweg
fault	breuklijn
fault	verschuiving
fault	breuk
fault line	breuklijn
feet	meter
feet	voeten
field	veld
foot	voet
foot of the crater	voet van de kraterwand
furnace	oven
gases	gassen
geothermal power	geothermische energie
giant ball	reusachtige bol
glaciers	gletsjers
golf	golfers
golf courses	golfbanen
golfers	golfers
ground	grond
ground	rotsen
ground	aarde
ground	ondergrond
harnesses	pakje
heat	hitte
heat	warmte
huge underground chambers	reusachtige ondergrondse kamers
incredible space shield	aardmagnetisch veld
inner core	binnenkern
land	land
landscapes	landschappen
lava	lava
lava	vulkaan
lava lake	lavameer
lakes	meren
life on earth	leven op aarde
limestone	kalksteen
light	licht
machine	machine
machine	activiteit
magnetite	magnetiet
magma	magma
magnetic field	magnetisch veld

magnetic field	veld
mantle	mantel
mantle	buitenmantel
metal	metaal
miles	kilometer
miles	meter
minerals	mineralen
movement	bewegingen
network	netwerk
northern coast	noordkust
oceans	oceanen
outer core	buitenkern
parking lots	parkeerterreinen
particular fault line	breuklijn
places	plaatsen
planet	aarde
planet	planeet
planet	oceanen
planet within a planet	planeet in een planeet
plates	platen
plate	plaat
population	mensen op aarde
power	energie
power	geothermische energie
procession	schouwspel
quake	aardbeving
result	resultaat
rim	rand
rocks	aardkorst
rock	gesteente
rock	rotsen
roof	dak
sample	staal
scientist	wetenschapper
scientists	wetenschappers
sea	bodem
shopping malls	shoppingcentra
side	kant
side of the fault	kant van de breuk
snow	sneeuwlandschap
solid rock	rotsen
source	bron
stadium	stadion
stations	centrales
streets	straten
sun	zon

surface	aardoppervlak
surface	grond
surface	oppervlakte
surface of the sun	oppervlakte van de zon
storm-tossed seas	woelig water
tank	watertank
team	team
team	klimwerk
towns	steden
turtle	schildpad
turtles	schildpadden
turtles	tank
types	groepen
tectonic plate	tektonische plaat
thermal energy	warmte uit de aarde
tectonic plates	tektonische platen
tectonic plates	platen
transponder system	transponder
underground chambers	ondergrondse kamers
vast network	uitgestrekt netwerk
virtual earth	virtuele aarde
volcano	krater
volcanoes	vulkanen
water	water
waterfront	waterkant
wall	borrelen
west	west

1b. The Earth Machine – Gold Standard glossary

Africa	#Afrika
age#	*tijd
Air	#lucht
Alaska	#Alaska
allow *** escape	#komen***vrij
Alps	#Alpen
America	#Amerika
Analyse	#onderzoek
Analysis	#onderzocht
Apple	#appel
apple’s skin	#schil van die appel
Aquifer	#aquifer

Aquifer	#ondergrondse grot
Arctic Circle	#noordpoolcirkel
Arctic Circle	#poolcirkel
Area	#gebied
Area	#landen
Area	#regio
Asia	#Azië
Atmosphere	#atmosfeer
atomic particles	geladen deeltjes
ball of *** lava	#bol***lava
ball of metal	#bol metaal
ball of superhot lava	#bol hete lava
Ball	#bal
Bay	#baai
Bay	#water van de baai
Beach	#strand
Bears	#Bears
bed of the lake	#bodem van het meer
Bed	#bodem
Biologist	#bioloog
Bleachers	#tribunes
blocks of concrete	#betonnen structuur
Boils	#borrelen
Bottom	#bodem
Bottom	#onderste laag
bubble over	#over * borrelen
building *** developments	#bouw- *** *projecten
building or farming developments	#bouw- of landbouwprojecten
Buitenkern	#outer core
California	#Californië
Celsius	#Celsius
centre of the earth	#kern van de aarde
centre of the earth	#middelpunt van de aarde
Centre	#kern
Chambers	#kamers
chemical composition	#chemische samenstelling
Chemical	#chemische
Chili	#Chili
China	#China
Churns	#pruttelen
Cities	#dorp
City	#stad
claustrophobic dive	#claustrofobie
coast of north and south America	#kustlijn van van Noord- en Zuid-Amerika
Coast	#kust*
Coast	#kustlijn

Collide	#botsen
column of superheated magma	#zuil erg heet magma
Column	#zuil
compasses	#kompassen
composition	#samenstelling
concrete	#betonnen
constant	#constant
continental plate	#continentale plaat
continents	#continenten
Cooled	#afgekoeld
Cools	#gaat afkoelen
Cools	#stolt
Core	#*kern
Country	#land
countryside	#landschap
cracks in the crust	#scheuren in de aardkorst
Cracks	#scheuren
crater wall	#kraterwand
Crater	#krater
Create	#ontstond
Creep	#schuiven
crunch together	#langs elkaar schuren
Crunch	#schuren
Crush	#hoge druk
crust of solid rock	#korst van vast gesteente
Crust	#*korst
Crust	#aardkorst
Current	#stroom
Cycle	#cyclus
deflected away	#afgebogen worden
degrees celcius	#graden celsius
degrees celcius	#graden
Degrees	#graden
Democratic Republic of the Congo	#Democratische Republiek Congo
Descent	#naar beneden
diamonds	#diamant
distances	#eind
Divers	#duikers
drinking water	#drinkwater
driven by	#aangedreven
earth machine	#aarde *** machine
earth machine	#machine
Earth	#aarde
Earth	#wereld
earth’s crust	#aardkorst
earth’s inner core	#binnenkern van de aarde

earth’s magnetic field	#aardmagnetisch veld
earth’s surface	#aardoppervlak
earth’s tectonic plates	#tektonische platen
earthquake activity	#aardbevingen
earthquake proof	#bestand tegen aardbevingen
Earthquake	#aarbeving
earth’s crust	#aardkorst
earth's inner core	#binnenkern van de aarde
earth’s interior	#diep vanbinnen
earth's magnetic field	#magnetisch veld
earth’s surface	#aardoppervlak
earth's tectonic plates	#tektonische platen
earth’s volcanoes	#onze vulkanen
Earths	#aard*
earth's	#van de aarde
eartquake activity	#aarbevingen
East	#oost
Edge	#rand
electric cables	#elektriciteitskabels
electrical current	#elektrische stroom
Electricity	#elektriciteit
Elements	#elementen
Empire State Building	#Eiffeltoren
Energy	#energie
Engine	#motor
Engineer	#technicus
eroded it to create	#En zo ontstond
Erupted	#kwam *** tot een uitbarsting
Erupted	#steeg
Erupted	#uitbarsting
Erupts	#knalt *** dwars door *** heen
eternal snows	#eeuwige sneeuw
Everest	#Everest
Exit	#uitweg
Experiment	#experiment
Extends	#loopt *** door
farming developments	#landbouwprojecten
fault line	#breuk
fault line	#breuklijn
fault needs to shift	#Plaatverschuivingen
fault system	#breuklijn
Fault	#breuk
Fault	#breuklijn
Fault	#verschuiving
Features	#verschijnselen
Feet	#meter

Field	#veld
fiery ball of solid metal	#gloeiende bal vast metaal
Floated	#dreven
Floor	#bodem
Florida	#Florida
Floridian Aquifer	#Floridan Aquifer
Flowing	#stromende
foot of the crater	#voet van de kraterwand
Foot	#meter
Foot	#voet
Forces	#krachten
free-floating blocks of concrete	#vrij bewegende betonnen structuur
Fumes	#*dampen
Furnace	#oven
Gallons	#liter
Gas pipes, water mains	#gas- en waterleidingen
Gases	#gassen
generated	#opgewekt
generates	#doet***ontstaan
generator	#generator
Geochemist	#geochemicus
geological features	#geologische verschijnselen
geological	#geologische
geologist	#geoloog
geothermal power	#geothermische elektriciteit
give away	#loskomen
Glacier Bay National Park	#Glacier Bay National Park
Glaciers	#gletsjers
goal post	#goal
going up	#stijgen
Golden Bears stadium	#Memorial Stadium
golden bears	#Golden Bears, het footballteam van de Universiteit van Californië
golf course	golfbaan
granite quarry	#granietgroeve
Granite rock	#graniet
Granite	#graniet
Ground	#aarde
Ground	#berg
Ground	#grond
Ground	#ondergrond
Ground	#rotsen
H20	#H20
hacked out	#gewonnen
Hayward fault	#Haywardbreuk
Heat	#hitte

Heat	#hoge *** temperatuur
Heat	#verwarmen
Heat	#warmte
heats up	#warmt op
Highest	#hoogste
Himalayas	#Himalaya
human being	#mens
Hydrogen	#waterdamp
ice age	#ijstijd
Ice	#ijs
ice-free	#ijsvrij
Iceland	#Ijsland
Icelanders	#Ijslanders
igneous rock	#stollingsgesteente
Igneous	#stollingsgesteente
Inch	#centimeter
Inch	#millimeter
Incline	#afdaling
India	#India
inner core	#binnenkern
interior of the planet	#diepste van de aarde
Japan	#Japan
kind of rock	#soort gesteente
lake of lava	#lavameer
Lake	#*meer
lake's surface	#oppervlakte van het meer
land mass	#landmassa
Land	#land*
Landscape	#landschap
lava lake	#lavameer
Lava	#lava
Lava	#uitbarstingen
Ledges	#terrassen
level ground	#beneden
Life	#leven
Limestone	#kalksteen
limitless * source of energy	#onuitputtelijke * bron van energie
line of the fault	#breuklijn
little marine creature	#schildpad
Local	#uit de streek
Location	#traject
loggerhead turtles	#onechte karetschildpadden
Los Angeles	#Los Angeles
Machine	#machine
machinery of the earth	#aarde
Magma	#magma

magnetic field	#aardmagnetisch veld
magnetic field	#magnetisch veld
magnetic field	#veld
magnetic shield	#magnetisch veld
magnetic	#*magnetisch
magnetite	#magnetiet
mantle plume	#mantelpluim
Mantle	#aardmantel
Mantle	#binnenmantel
Mantle	#buitenmantel
Mantle	#mantel
Maps	#registreert
Marble	#marmer
marine creature	#schildpad
massive expanse of water	#oceaan
Matrix	#netwerken
Melted	#smelten
Memorial Stadium	#Memorial Stadium
Metal	#metaal
Metal	#metalen
metamorphic rock	#metamorf gesteente
miles above sea level	#meter boven zeeniveau
Miles	#kilometer
Miles	#meter
Mineral	#mineraal
molten metal	#gesmolten metaal
molten metal	#vloeibaar metaal
molten rock	#gesmolten gesteente
molten rock	#lava
molten rock	#vloeibare gesteente
monitored	#houdt *** in de gaten
Moon	#maan
Motion	#beweging
Mount Everest	#Mount Everest
Mount Fairweather golf course	#Mount Fairweather Golfclub
Mount Nyriagongo	#Nyiragongo
mountain range	#bergketen
mountains	#bergen
mountains	#gebergte
Move	#bewegen
Move	#verschuiven
Moved	#opgeschoven
movement of the crust	#bewegingen van de aardkorst
movement	#beweging
movement	#schok
National park	#National Park

natural structures	#natuurlijke structuren
Network	#netwerk
never-ending movement of the land	#continue bewegingen van het land
New York State	#Amerikaanse staat New York
New Zealand	#Nieuw-Zeeland
Night	#nacht
Niyrangongo	#Niyrangongo
north America	Noord-Amerika
North American tectonic plate	#Noord-Amerikaanse plaat
North American	#Noord-Amerikaanse
north and south America	#Noord- en Zuid-Amerika
North Atlantic Ocean	#Noord-Atlantische Oceaan
North Carolina	#Amerikaanse staat North Carolina
North	#noord
northern California	#Noord-Californië
northern coast of California	#noordkust van Californië
northern coast	#noordkust
Northern Lights	#noorderlicht
Northern lights	#poollicht
Ocean	#oceaan
on the move	#in beweging
open faced granite Quarry	#open granietgroeve
outer core	#buitenkern
outer surface	#buitenste laag
Oxygen	#Zuurstof
oxygen#zuurstofflessen#3	#Zuurstof
pacific Rim	#pacific Rim
pacific Rim	#pacific Rim,landen rondom de Stille Oceaan
Pacific	#Stille Oceaan
Park	#Park
parking lots	#parkeerterreinen
Particles	#deeltjes
Peak	#berg
Planet Earth	#planeet Aarde
Planet	#aarde
Planet	#planeet
Planet	#wereld
planet’s generator	#generator van onze planeet
Plate	#plaat
Plates	#platen
Plume	#*pluim
plumes of heat	#pluimen
Plumes	#pluimen
Polished	#gepolijst
Pollute	#vervuilen
Population	#mensen

Power stations	#Geothermische centrales
Power	#elektriciteit
Power	#energie
Power	#geothermische energie
Power	#kracht
powered by	#aangedreven door
pressure	#druk
pressure	#spanning
Process	#proces
procession of lights	#schouwspel van kleuren
procession	#schouwspel
Proof	#bestand tegen
pushing together	#zijn *** tegen elkaar aan het duwen
Pushing	#duwen
put to work	#aandrijfkracht
Quake	#aardbeving
Quakes	#bevingen
Quarry	#*groeve
radar images	#radarbeelden
Radiates	#straalt * uit
radio engineer	#technicus
Ranger	#parkwachter
Records	#legt *** vast
resource	#bron
Reykjavik	#Reykjavik
Rift Valley	#Grote Riftvallei
Rift Valley	#Riftvallei
Rim	#rand
Rises	#stijgt
Rivers	#rivieren
Rivers	#zeeën
rock ***metamorphic	#metamorf gesteente
Rock	#gesteente
Rock	#graniet
Rock	#rotsen
Rocks	#gesteente
Rocks	#rotsen
roof of the aquifer	#dak van de aquifer
Route	#route
Russia	#Rusland
Sample	#staal
Samples	#staal
San Andreas Fault	#San Andreasbreuk
San Francisco	#San Francisco
satellite maps	#satellietkaarten
Satellite	#satelliet*

Scientist	#collega
Scientist	#wetenschapper
sea floor	#bodem van de oceanen
sea level	#zeeniveau
Sea	#oceanen
Seabed	#zeebodem
Seas	#water
Seawater	#zeewater
Section	#deel
sedimentary rocks	#sedimentair gesteente
Sedimentary	#sedimentair
seeped down	#door *** sijpelen
Seeps	#sijpelt
separate free-floating blocks of concrete	#aparte, vrij bewegende betonnen structuur
shaft of light	#lichtstraal
sheet of ice	#laag ijs
Sheet	#laag
Shield	#veld
Shift	#Plaatverschuivingen
Shifted	#verschoof
shopping malls	#shoppingcentra
Shore	#water
shores of North Carolina	#kust van North Carolina
Shores	#kust
Sidewalk	#stoeptegels
Skies	#luchten
Skies	#luchten…
slide by one another	#schuiven *** langs elkaar
Slide	#schuiven
Slip	#schuren
Snow	#sneeuwlandschap
Snows	#sneeuw
solid metal	#vast metaal
solid rock	#rotsen
solid rock	#vast gesteente
Solid	#vast
solidified into rock	#gestold
Solidified	#gestold
sonar *** images	#sonar- *** beelden
sonar and radar images	#sonar- en radarbeelden
sonar images	#sonarbeelden
source of *** power	#bron van ***kracht
source of energy	#bron van energie
Source	#bron
south America	#Zuid-Amerika
south coast of California	#zuidkust van Californië

south coast	#zuidkust
south to the northern coast of California	#zuid- tot de noordkust van Californië
South	#zuid
south*** America	#Zuid-Amerika
southern California	#Zuid-Californië
southern states	#zuidelijke Amerikaanse staten
space shield	#aardmagnetisch veld
Space	#ruimte
special suit	#speciaal pak
Sphere	#bol
spilling over	#vloeien * over
Split	#opgesplitst
Splits	#scheuren
springing *** up	#veert *** op
square miles	#vierkante kilometer
Stadium	#stadion
Stadium	#station
Stands	#muren
States	#staten
Stations	#centrales
Steam	#stoom
storm-tossed seas	#woelig water
Stresses	#spanningen
structures	#structuren
subterranean rivers	#ondergrondse rivieren
subterranean world	#onderaardse wereld
subterranean	#ondergrondse
subway tunnels	#metro’s
Suit	#pak
Sulphurous fumes	#zwaveldampen
Sulphurous	#zwavel*
Sun	#zon
sunshine state	#zonovergoten staat
super-continent	#supercontinent
super-heated molten rock	#gloeiend hete lava
surface area	#oppervlakte
surface of the planet	#aardkorst
surface of the planet	#aardoppervlak
surface of the planet	#oppervlakte van de aarde
surface of the sun	#oppervlakte van de zon
surface of the world	#aardoppervlak
Surface	#*korst
Surface	#*oppervlak
Surface	#aarde
Surface	#aardoppervlak
Surface	#boven de grond

surface#grond	#boven de grond
Surface	#laag
Surface	#oppervlak
Surface	#oppervlakte
surges over	#spuit * over
Survey	#bestudeert
Surveyed	#in kaart gebracht
sweeping glaciers	#uitgestrekte gletsjers
Sweeping	#uitgestrekte
Tank	#tank
tectonic plate	#plaat
tectonic plate	#tektonische plaat
tectonic plates#tektonische platen	#tektonische plaat
Temperature	#temperatuur
thermal energy	#warmte uit de aarde
thin crust of solid rock	#dunne korst van vast gesteente
Tons	#ton
tore the land apart	#scheurde het land * open
Towers	#torent
Towns	#steden
transponder system	#transponder
Tsunami	#tsunami
Turtle	#schildpad
Turtles	#schildpadden
types of rock	#groepen gesteente
types of rock	#soorten gesteenten
underground aquifers	#aquifers
underground chambers of magma	#ondergrondse kamers vol magma
underground chambers	#ondergrondse kamers
Underground	#ondergrondse
Underwater	#blank
University of California	#Universiteit van Californië
University of California, Berkeley	#Universiteit van Californië
University of Rochester in New York State	#Universiteit van Rochester in de Amerikaanse staat New York
Unmapped	#ongemerkt
upwelling of intense heat	#enorme stroming van opwaartse hitte
Valleys	#valleien
Vast	#uitgestrekt
Verschuiving	#is *** moving
very bottom of the mantle	#onderste laag van de binnenmantel
very bottome laag	#onderst
virtual earth	#virtuele aarde
virtual planet earth	#virtuele planeet Aarde
Virtual	#virtueel
volcanic eruptions	#vulkaanuitbarstingen

Volcanic	#vulkaan*
Volcano	#de Nyiragongo
volcano's crater	#krater
wall of rock	#rotswand
Wall	#*wand
water mains	#waterleidingen
water tank	#watertank
Water	#oceaan
Water	#water
waterfront	#waterkant
Waves	#golven
Weight	#gewicht
West	#west
Winter	#winter
World	#aard*
World	#wereld

2a. Madagascar , The Island of Marvels – TExSIS glossary

adolescent	puber
ant	mier
animal	dier
animals	dieren
bent	scheefgegroeid
bent by the wind	scheefgegroeid door de wind
bird	vogel
bits	stukjes
bone	getuigen
birds	vogels
block	kalksteenplateau
branch	tak
caves	grotten
chameleon	kameleon
chameleons	loop
character	karakter
cobwebs	spinrag
creature	schepsel
dark	duisternis
design	ontwerp
distant land	ver land
drought	droogte
dryness	droogte
drongo	kuifdrongo
eyes	ogen

earth	aarde
earthquakes	aardschokken
extraordinary story	fabelachtig verhaal
far south	zuiden
farthest southerly point	uiterste zuidpunt
female	wijfje
female	partner
female	koppeltje
flower	bloem
floor	grond
forest	boeg
forest	zwammenbos
forested slopes	beboste hellingen
fossa	fretkat
fragments	eierschalen
front	voorkant
gap	begane grond
giraffe	giraffenkever
great block	reusachtig kalksteenplateau
grass	gras
ground	grond
hands	grijphanden
heart	ondergrond
heat	hitte
hedgehog	egel
hole	gat
immense lake	reusachtig zoutmeer
insects	insecten
indri	indri
intruder	indringer
intense dryness	extreme droogte
island	eiland
journey	tocht
kilometres	kilometer
lake	zoutmeer
land	land
landscape	eiland
leaf	blad
leaf litter	helling
leaves	bladeren
lemurs	maki's
lemurs	kroonmaki's
legs	poten
lemur	maki
limestone	bodem
loser	verliezer

male	mannetje
metres	meter
nest	nest
north	noorden
north of the island	noorden van het eiland
notches	randen
nutrients	voedsel
oasis	oase
outsiders	buitenstaanders
pioneering castaways	gestrande pioniers
place	plek
plants	planten
past	geologisch verleden
pursuit	zoek
rain	regen
rain	neerslag
rainforest floor	begane grond
red female	rosse wijfje
red male	ros mannetje
ringtails	ringstaartmaki's
river forests	rivierbossen
rock	gesteente
roots	wortels
rivers	rivieren
set of animals	clubje dieren
shell	huisje
shelter	schaduw
sifakas	sifaka's
shores	kust
slopes	hellingen
sources	snippers bos bulken
southern river system	rivierenstelsel
species	soorten
spider	spin
spiny trees	doornige bomen
spiny woodlands	vol doorn
spine	bergketen
sunbird	honingzuiger
survivors	primitieve schepsels
story	verhaal
strand	draad
tenrec	tenrek
territory	weg
territory holder wins	eigenaar
thousands	scherven
trees	bomen

toes	tenen
tree	boom
tropics	tropen
tropical island	tropische eiland
underground rivers	ondergrondse rivieren
variation	verschillen
vast hole	immens gat
vegetation	planten
volcanoes	vulkanen
water	water
waves	golven
west	westen
western side	westelijke helft
white male	witte mannetje
wind	wind
world's rarest carnivores	zeldzaamste roofdieren op aarde

2b. Madagascar, The Island of Marvels – Gold Standard glossary

adapt to	#wennen aan
Adapt	#wennen
Adaptable	#passen zich enorm goed aan
Adapted	#aangepast
Adapted	#flexibel
Adolescent	#puber
Adult	#volgroeid
Africa	#Afrika
African predators	#roofdieren van Afrika
African	#van Afrika
Ancestors	#voorouders
ancient creatures	#primitieve schepsels
Animal	#dier
Animals	#dieren
Ant	#mier
Ant	#mier
Antarctica	#Zuidpool
are under threat	#dreigen
Area	#streek
arid world	#wereld zonder wat er
Arid	#zonder water
Ashore	#op het strand
astonishing creature	#olifantsvogel
Babies	#jongen
Babies	#kleintjes
Baby	#jong
back legs	#achterpoten

Baobab	#apenbroodboom
Baobab	#baobab
Baobab	#baobab, de apenbroodboom
bare rock	#kale rots
Beach	#kust
Behaviour	#gedrag
Bird	#vogel
bits of bone	#stukjes bot
bloated trunks	#dikke stam
block of limestone	#kalksteenplateau
Bone	#bot
Branches	#takken
Breed	#broeden
Breeding territory	#broedplekken
Brother	#broer
Bush	#struik
Canyons	#ravijnen
Carnivores	#roofdieren
Carved	#uitgesleten
carving holes into the limestone	#gaten uit de kalksteen slijt
carving holes	#gaten *** slijt
Carving	#slijt
cast adrift	#sloeg op drift
cast away	#in afzondering leeft
Caverns	#grotten
Caves	#grotten
Centre	#centrum
Chameleon	#kameleon
Childhood	#kindertijd
Children	#jongen
Cliffs	#klippen
Climate	#klimaat
Climbing	#klimmers
Climbs	#kan *** klimmen
Cobwebs	#spinrag
Collapsed	#ingezakt
collection of wildlife	#collectie dieren
Competition	#concurrentie
Conditions	#klimaat
Continents	#continenten
Coua	#coua
could well be	#dreigen
Couple	#koppeltje
Cracks	#spleten
Creases	#kreukjes
Creature	#olifantsvogel

Creature	#schepsel
Crossing	#over te steken
crowned lemurs	#kroonmaki's
Curiosities	#rare kwanten
Curl	#opgerolde tip
Curling	#rollen
curls up	#rolt *** op
Danger	#indringer
Descendents	#stammen af
Desiccation	#Uitdroging
Develop	#variatie te brengen
Developed	#ontwikkeld
Dier	#inhabitants
dinosaur eggs	#die van een dinosaurus
Dinosaurus	#dinosaurus
Disappeared	#verdween
Disappearing	#verdwijnen
Discover	#beseffen
dissolved away	#het werk van
diversified into	#evolueerden * tot
Diversified	#evolueerden
diversity of life	#diversiteit aan leven
diversity of wildlife	#collectie dieren
Diversity	#collectie
Diversity	#diversiteit
dog-like	#honden
Dotted	#bezaaid
drenched in rain	#verzuipt in de regen
Drenched	#druipnat
driest times	#droogste periodes
Drongo	#drongo
Drongo	#kuifdrongo
Drought	#droogte
Drowned	#lag *** onder de zeespiegel
dry landscape	#dorre milieu
Dryness	#droogte
Earth	#aarde
Earthquakes	#aardschokken
east coast	#oostkust
East	#oosten
eastern side	#oostkant
Edges	#oevers
egg fragments	#kapotte eierschalen
egg shells	#eierschalen
Egg	#ei
elephant bird	#olifantsvogel

Endangered	#bedreigd
Evaporating	#verdampt
evenly matched	#aan elkaar gewaagd
Evolutionary	#evolutionaire
Evolved	#ontstaan
Evolving	#ontwikkelde zich
expectant fathers	#toekomstige vader
extinct in the wild	#voorgoed te verdwijnen
extra long neck	#lange nek
Eyes	#ogen
family group	#gezinnetje
farthest southerly point	#uiterste zuidpunt
Fathers	#vader
Feed	#eten
female weevil	#wijfje
Female	#partner
Female	#wijfje
Fertile	#loops
few final folds	#paar keer vouwen
Fighting	#vecht
Firewood	#brandhout
Fish	#vissen
Flamingos	#Flamingo's
Floats	#drijven
Floor	#grond
Flower	#bloem
Flows	#loopt
flycatcher couple	#koppeltje
Flycatchers	#paradijsmonarchen
Food	#voedsel
forest floor	#grond
forest floors	#begane grond
forest of toadstools	#zwammenbos
forest pockets	#bosjes
Forest	#bomen
Forest	#bos
forested slopes	#beboste hellingen
Forested	#bebost
Forests	#regenwouden
Fossa	#fretkat
fragments of egg shells	#scherven van eierschalen
fresh water	#zoet water
fringe of forest	#smalle strook bos
fruiting trees	#fruitbomen
Fused	#paarsgewijs aaneengegroeid
Gallop	#huppelen

geological history	#geologisch verleden
Geological	#geologisch
getting drier	#uitdrogen
giant fig	#wurgvijg
giant mongoose	#fretkat
giant pinnacles	#vlijmscherpe spitsen en pieken
giant world	#reuzenwereld
giraffe necked weevil	#giraffenkever
giving birth	#werpen
gnarled and twisted spiny woodlands	#knoestige dwergstruiken vol doorn
Grandidier’s vontsira	#grandidier s
Grandidier’s vontsira	#grandidoer mangoeste
grasping hands and feet	#typische grijphanden
grasping hands	#grijphanden
Grass	#gras
Grasses	#grassen
great slab of land	#oercontinent Gondwana
Greater flamingos	#Flamingo's
Green	#groen
Grijze	#verreaux
grip like tongs	#geeft meer greep
Grip	#greep
Ground	#grond
Habitat	#leefgebied
Habits	#prioriteiten
has adapted to live here	#heeft van dit ontoegankelijke terrein zijn thuis gemaakt
Hatch	#uitkomen
have diversified to an astonishing degree	#vormen nu een heel divers gezelschap
heart of Madagascar	#ondergrond van Madagaskar
Heat	#hitte
Hedgehog	#egel
hind legs	#achterpoten
History	#verleden
Holes	#gaten
huge forests populated with strange, bulging trees	#immense bossen met bizarre, opgezwollen bomen
Humans	#mensen
Hunted	#gejaagd
India	#India
Individual	#dier
Indri	#indri
Insects	#insecten
Intruder	#indringer
Island	#eiland
island's top predator	#fretkat

isolated forests	#snippers bos
Isolated	#afgesneden
Isolated	#snippers
Isolation	#geïsoleerde ligging
Isolation	#isolement
jumping skills	#springtechnieken
Jungle	#jungle
Lac Alaotra reed lemur	#Alaotra-bamboemaki
Lac Alaotra	#Alaotra-meer
lake of salt	#zoutmeer
Lake	#meer
Land	#eiland
Land	#land
Land	#vasteland
landscape of scrub and spines	steppelandschap van dor en doornig struikgewas
Landscape	#landschap
Landscape	#milieu
Landscape	#steppelandschap
Landscape	#stuk van het eiland
layers of limestone	#lagen kalk
Layers	#lagen
leaf litter	#rottende bladeren
leaf nests	#bladeren
leaf roll	#stengel
Leaf	#blad
Leaf	#bladeren
leaf’s ribs	#randen van het blad
leaf’s veins	#nerven van de bladeren
Leaping	#springen
leat litter	#rottende bladeren
Leaves	#bladeren
Legs	#poten
Lemur	#maki
Lemurs	#maki's
Leopards	#luipaarden
life in the trees	#in de bomen leeft
life in the trees	#leven in de bomen
Life	#leven
Limestone	#bodem
Limestone	#kalk
Limestone	#kalksteen
Lions	#leeuwen
little nocturnal mammal	#nachtdiertje
little oases of forest	#groepjes bomen, kleine oases
little sister	#zusje

long roots	#lange wortels
lush jungle	#weelderige jungle
Madagascar	#eiland
Madagascar	#Madagaskar
Madagascar’s biggest lake	#grootste meer van Madagaskar
Madagascar’s equivalent of hedgehogs	#plaatselijke equivalent van onze egel
Madagascar’s centre	#centrum van Madagaskar
Madagascar's wild landscapes and speciesMadagaskar#EN_LOC	#oorspronkelijke landschappen en diersoorten van
Madagascar’s wildlife	#dieren op Madagaskar
male panther chameleon	#panterkameleon
Male	#mannetje
Male	#vrijer
Mammal	#zoogdieren
Mammals	#zoogdieren
Mate	#paren
Mate	#partner
Material	#nestmateriaal
mating season	#paarseizoen
Mating	#aan het paren
Mats	#matten
Millipede	#duizendpoot
Miniatures	#minuscuul klein
Moles	#mollen
Mongoose	#mangoeste
Mother	#Moeder
mouse-sized	#zo groot als een muis
Neck	#nek
Nectar	#nectar
nest sites	#nestplaatsen
Nest	#nest
nocturnal creatures	#nachtdieren
nocturnal mammal	#nachtdiertje
nocturnal, mouse-sized creatures	#nachtdieren amper zo groot als een muis
north of the island	#noorden van het eiland
North	#noorden
Notches	#inkepingen
Nutrients	#voedsel
Oases	#oases
Oasis	#oase
observed in the wild	#waargenomen
Ocean	#oceaan
Ocean	#zee
Oddities	#bizarre schepsels
Oddities	#curiosa
Oddity	#rariteiten

Origins	#roots
orphaned chip of land	#verdwaald restje land
panther chameleon	#panterkameleon
paradise flycatcher	vlijmscherpe randen
Parasitise	#parasieten
parched and sandy wilderness	#verschroeide zandvlakte
Partner	#partner
Past	#verleden
patch of reed	#rietveld
patch of reeds	#rietveld
Peaks	#pieken
Pigment	#pigment
Pinnacles	#pieken
Pinnacles	#spitsen
pioneering castaways	#gestrande pioniers
Plants	#planten
Plants	#struik
Plateau	#plateau
Pollen	#stuifmeel
Pollinators	#bestuivers
Practice	#oefening
Predators	#roofdieren
Primate	#primaten
Primates	#apen
Primates	vlijmscherpe randen
primitive mammals	#primitieve zoogdieren
Primitive	#primitieve
pygmy chameleon	#dwergkameleon
radiated tortoises	#stralenschildpad
radiated tortoises	#stralenschildpaden
rain shadow	#droogte
Rain	#neerslag
Rain	#regen
rainforest edge	#rand van het regenwoud
rainforest floor	#begane grond
rainforest floor	#grond
Rainforest	#regenwoud
Rainforests	#jungle
Rainforests	#regenwoud
rare specialists	#dieren *** die enorm gespecialiseerd en zeldzaam zijn
Rare	#buitenbeentjes
Rare	#zeldzaam
razor sharp blades of stone	vlijmscherpe randen
red feathers	#ros
reed bed	#rietveld

reed beds	#rietvelden
reed beds	#rietveldjes
Reeds	#riethalmen
refuge from the heat	#beschutting tegen de hitte
Refuge	#beschutting
Relation	#familie
remote landscape	#desolate landschap
Reptile	#reptielen
Ribs	#randen
rich forest floors	#begane grond
Ringtailed lemurs	#ringstaartmaki's
Ringtails	#ringstaartmaki’s
ripped from	#losgescheurd van
river forests	#rivierbossen
river system	#rivierenstelsel
Rivers	#rivieren
Roamed	#zwierf
Rock	#gesteente
Rock	#rots
Rolling	#vouwen
Roots	#wortels
rumbles with geological activity	#rommelt
salt lake	#zoutmeer
Sand	#zand
sandy wilderness	#zandvlakte
scrub and spines#dor en doornig	#beschutting struikgewas
Scrub	#dor
Scrub	#struikgewas
Scrublands	#Madagaskar
Seabed	#land
Shadows	#vlekken licht en schaduw
sheer cliffs	#loodrechte muur
Shell	#huisje
Shell	#schild
Shelter	#schaduw
Shores	#kust
Shrews	#spitsmuizen
Side	#helft
Sifakas	#sifaka's
Silk	#zijdedraden
Sister	#zusje
skill and practice	#techniek en oefening
Skill	#techniek
Slopes	#helling
slumbering volcanoes	#sluimerende vulkanen
snail shell	#slakkenhuis

snipping through	#knipt *** door
snips *** off	#bijt *** door
soft leaf	#zacht blad
Sources	#bulken van
South	#zuid
South	#zuiden
southerly point	#zuidpunt
southern scrublands	#zuiden
Southern	#zuiden
Species	#diersoorten
Species	#soorten
Speckled	#bezaaid
Spider	#spin
spine of mountains	#bergketen
Spines	#doornig
spiny scrublands	#struikbosjes
spiny treesen	#doornige bom
spiny woodlands	#dwergstruiken vol doorn
Spiny	#doornige
steep volcanic slopes	#helling
Stem	#stam
Strand	#draad
Stripy	#gestreept
struggle to live anywhere else	elders niet zouden aarden
Sunbird	#honingzuiger
Surefooted	#niet uit haar evenwicht te krijgen
Surface	#oppervlak
Survive	#leven
Swim	#zwemmen
Tactic	#tactiek
tail plumes	#staartveren
Technique	#techniek
Tenrec	#tenrek
Tenrecs	#Tenreks
territory holder	#eigenaar
Territory	#beschikbare ruimte
Territory	#ruimte
thirsty roots	#dorstige wortels
Thrive	#gekoloniseerd
Times	#periodes
Tiptoes	#kruipt voetje voor voetje
Toes	#tenen
tough grass	#taai gras
Tree	#boom
Trees	#bomen
Trees	#bomen

tropical island	#tropische eiland
Tropics	#tropen
Trunk	#boomstam
Trunks	#stam
Tsingy	#Tsingy
Types	#soorten
Uncarina#uncarina	#soorten
underground rivers	#onderaardse rivieren
underground rivers	#ondergrondse rivieren
Underground	#onder de grond
Underneath	#ondergrond
Underwater	#op de bodem
uplifted rock	#omhooggestuwd gesteente
Vanish	#verdwijnen
Variation	#sterk wisselende
varied landscapes	#variatie in het landschap
vast hole right in these central uplands	#immens gat
Vegetation	#planten
Veins	#nerven
Verreaux's coua	#grijze coua
volcanic fire	#vulkanische activiteit
volcanic forest	#vulkaanbos
volcanic slopes	#helling
Volcanoes	#vulkanen
Wafts	#zwaait
walls of rock	#muren
was * abundant	#floreerden
washed in	#aanspoelden
washed in	#belandden hier als drenkelingen
Water	#water
Water	#wind en regen
Waves	#golven
West	#westen
western side	#westelijke helf
western side	#westkant
wide plateau of uplifted rock	uitgestrekt plateau van omhooggestuwd gesteente
wild dogs	#hyena's
wild landscapes	#oorspronkelijke landschappen
Wildlife	#dieren
Wind	#wind
Wind	#windvlaag
windswept beach	#kust
Woodlands	#dwergstruiken
Youngsters	#jongen

3a. The Secret World of Pain – TExSIS glossary

artist	kunstenares
attention	aandacht
babies	babys
brain	hersens
brain stem	hersenstam
changes	schade
child	pijnbestrijding
child pain management	pijnbestrijding bij kinderen
emotions	emoties
events	kleinkinderen
feel	ervaart
full-term babies	voldragen babys
function	functie
genes	genen
healing	ongeluk
heat	hitte
impact	pijn
influence	pijn
injury	genezing
lives	leven
low pain stimulus	lichte pijnschok
mind	hersens
motor cortex	motorische schors
neurons	neuronen
pain	pijn
pain	pijnmechanisme
pain	pijnsysteem
pain signals	pijnsignalen
pain stimulus	pijnschok
pain system	pijnsysteem
persistent pain	constant pijn
physiotherapy	fysiotherapie
pulses	impulsen
premature babies	premature babys
problem with heat	probleem met hitte
process	proces
scientists	wetenschappers
signals	pijnsignalen
team	team
treatment	behandeling

3b. The Secret World of Pain – Gold Standard glossary

advanced scanners	#scanners
Affect	#beïnvloeden

Affects	#bepaalt
Affects	#treft
Alter	#veranderen
Analysed	#analyseren
are defective	#het niet doen
area of the brain	#hersenstam
Areas of * brain	#hersengebieden
areas of *** brain	#hersengebieden
Areas of *** brain	#hersengebieden
Arm	#arm
Attention	#aandacht
attentional system	#aandacht
Babies	#baby's
bizarre condition	#afwijking
blocking the pain signals	#pijnsignalen tegen * houden
Body	#hele lichaam
Body	#lichaam
brain activity	#hersenactiviteit
brain scans	#hersenscans
brain stem	#hersenstam
Brain	#brein
Brain	#hersenen
Brain	#hersens
brain's attentional system	#aandacht
Burns	#brandwonden
chain of chemical reactions	#chemische reactie
chemical reactions	#chemische reactie
Childhood	#jeugdervaringen
chronic pain sufferers	#mensen die aan chronische pijn lijden
chronic pain	#chronische pijn
clinical trials	#kinderschoenen
clinically required	#voor ze naar huis mogen
collecting brain scans from	#neemt * hersenscans bij
Comparing	#vergelijken
complex mechanisms	#pijnmechanisme
Complex Regional Pain Syndrome	#complex regionaal pijnsyndroom
Connections	#verbindingen
countless chronic pain sufferers	#miljoenen mensen die aan chronische pijn lijden
countless experiments	#experimenten
curing it	#pijn behandelen
Curing	#behandelen
cut off	#amputeren
cutting edge treatment	#revolutionaire therapie
cutting edge treatments	#revolutionaire pijntherapie
Damaged	#beschadigd

Danger	#gevaar
decision-making	#besluitvorming
Develop	#ontwikkelen
Develop	#ontwikkeling
Develop	#zich ontwikkelen
Developing	#ontwikkeld
Developing	#richten *** op
Disappear	#gaat * over
Discovery	#ontdekking
DNA	#DNA
dramatic event	#dramatische gebeurtenis
Effects	#invloed
electrical signals	#elektrische impulsen
Emotion	#emotie
Emotions	#emoties
Environment	#omgeving
Event	#gebeurtenis
events in early childhood	#jeugdervaringen
Events	#eigenschap
excruciating pain	#ondraaglijke pijnen
experience pain	#pijn ervaren
Experience	#ervaren
experiences in the early days of life	#vroege pijnervaringen
Experiment	#experiment
Experiments	#experimenten
Explain	#verklaren
Explore	#zoekt uit
Extend	#geërfd
Fails	#faalt
faulty gene	#genetische afwijking
feel * pain	#voelt * pijn
feel pain	#pijn * ervaren
feel pain	#pijn ervaart
feel pain	#pijn voelen
feel pain	#pijnbeleving
feel pain	#voelen * pijn
Feel	#ervaart
Feel	#voelen
Feels	#lijdt
Feels	#voelt
Fingerprint	#vingerafdruk
fire magnetic pulses	#sturen * impulsen
fire pain signals	#stuurt * signalen
Flexible	#flexibel
full-term babies	#voldragen baby's
full-term baby	#baby

Function	#functie
GCSEs	#eindexamen
Gene	#gen
Generated	#ontstaat
Generations	#generaties
Genes	#genen
genetic fingerprint	#genetische vingerafdruk
genetic link	#genetisch bepaald
Geneticist	#geneticus
Genetics	#genen
group of scientists	#wetenschappers
harbour View Medical Center	#brandwondencentrum
have an impact on	#inwerkt op
Healed	#genezing
healing process	#genezingsproces
Heat	#hitte
heel prick	#hielprik
high pain stimulusx	#hielprik
how we feel pain	#pijnbeleving
Hurts	#pijn * doet
Impact	#impact
increased sensitivity to pain	#gevoeliger * voor pijn
Information	#informatie
Injury	#blessure
Injury	#letsel
Injury	#ongeluk
Insight	#blik
Investigate	#kijkt
Investigate	#uitzoeken
Investigating	#gaan * op zoek
Investigating	#onderzoekt
Investigating	#op zoek
is generated	#ontstaat
key region in the brain	#belangrijkste hersengebied
lasting injury	#blijvend letsel
life experiences	#levenservaringen
Life	#leven
Lives	#leven
low pain stimulus	#lichte pijnschok
magnetic pulses	#impulsen
magnetic pulses	#magnetische impulsen
magnetic stimulation	#magnetische stimulatie
manipulating that part of the brain's attentional system	#leiden *** aandacht af
medical health problems	#medische raadsels
medical science	#Wetenschappers

Messages	#boodschap
Messages	#signalen
Mind	#hersens
minor accident	#ongelukje
motor cortex	#motorische schors
Motor	#motorische
motorbike accident	#motorongeluk
Movement	#bewegingen
Mutations	#mutatie
Mysteries	#mysterie
natural painkillers	#natuurlijke pijnstillers
nerve cells	#zenuwcellen
nervous system	#zenuwbanen
Neurons	#neuronen
Neuroscientist	#neurowetenschapper
new treatment	#draaglijker
one of science’s greatest mysteries	#eeuwenoud medisch mysterie lichten
over-stimulated	#te veel gestimuleerd
pain experience	#pijnervaring
pain experts	#pijnexpert
pain is generated	#pijn ontstaat
pain mechanisms	#pijn
pain mechanisms	#pijnmechanisme
pain mechanisms	#pijnsysteem
pain pathways	#pijn
pain pathways	#pijnbanen
pain pathways	#zenuwbanen
pain protects	#Pijn beschermt
pain signals	#pijnsignalen
pain signals	#signalen
pain stimulus	#pijnschok
pain system	#pijn
pain system	#pijnmechanisme
pain system	#pijnsysteem
pain that persists	#pijn die aanhoudt
Pain	#pijn
Pain	#pijnbeleving
Pain	#pijnen
Pain	#pijnervaring
Pain	#pijnimpulsen
painful treatments	#pijnlijke behandelingen
Painful	#pijnlijk
Painful	#pijnlijke
Painkillers	#pijnstillers
part of * brain	#hersengebied
parts of * body	#lichaamsdelen

parts of her body	#lichaamsdelen
parts of the brain	#hersengebieden
patients with chronic pain	#patiënten met chronische pijn
Patients	#patiënten
perception of pain	#pijnbeleving
persistent pain	#pijn
Persists	#aanhoudt
physical pain	#lichamelijke pijn
Physiotherapy	#fysiotherapie
premature babies	#premature baby's
Problem	#probleem
Procedure	#pijnlijke prik
Process	#proces
Process	#stuurt
Process	#verwerken
protects and alerts	#beschermt
Protects	#beschermt
Pulses	#impulsen
Purpose	#nut
radical new therapies	#therapieën
receiving * pain signals	#krijgt * pijnsignalen
Recorded	#registreren
Recovery	#hersteld
reduce * pain	#pijn afnemen
reduce her pain	#haar pijn afnemen
region in the brain	#hersengebied
regions of the brain	#hersengebieden
Register	#registreert
Regulates	#regelt
Rehabilitation	#behandeling
relief from *** pain	#wil van *** pijn af
relieve *** pain	#pijn verlichten
Relieve	#verlichten
Research	#fenomeen
Research	#onderzoek
Results	#resultaten
reverse the changes	#schade * herstellen
Reverse	#herstellen
Saved	#gered
Scanners	#scanners
science's * mysteries	#medisch mysterie
Scientists	#wetenschap
Scientists	#wetenschappers
SCN9a gene	#gen SCN9A
SCN9a gene	#SCN9A
SCN9a	#gen SCN9A

SCN9a	#SCN9A
second * degree burns	#tweedegraads brandwonden
second and third degree burns	#tweede- en derdegraads brandwonden
sensation of pain	#pijnervaringen
sense of pain	#pijnbanen
Senses	#zintuigen
sensory neurons	#zintuiglijke neuronen
Signals	#impulsen
Signals	#signalen
simple injury	#lichte blessure
situations of extreme survival	#extreme situaties
Situations	#situaties
Skin	#littekenweefsel
source of pain	#plaats waar pijn ontstaat
Stimulated	#gestimuleerd
Stimulation	#gestimuleerd
Stimulation	#stimulatie
Stroke	#beroerte
Subjective	#subjectief
suffer ***pain	#lijden*** pijnen
suffer excruciating pain	#lijden * ondraaglijke pijnen
suffer second and third degree burns	#liep tweede- en derdegraads brandwonden op
Suffer	#lijden
suffered second and third degree burns	#liep tweede- en derdegraads brandwonden op
Suffered	#liep * op
Sufferers	#mensen die * lijden
suffering from persistent pain	#heeft constant pijn
Suffering	#ellende
suffers from	#lijdt aan
survival mechanism	#overlevingsmechanisme
Survival	#ons leven redden
Therapies	#therapieën
third degree burns	#derdegraads brandwonden
Touch	#aanraking
Trait	#eigenschap
trans-cranial magnetic stimulation	#transcraniële magnetische stimulatie
Transform	#veranderen
transmit messages	#signalen doorsturen
Transmit	#doorsturen
Transmit	#sturen
transmitted messages to	#zonden een boodschap naar
Transmitted	#zonden
Treat	#nieuwe behandelingen
Treatment	#behandeling

Treatment	#therapie
Treatments	#behandelingen
Treatments	#pijntherapie
turn * down	#blokkeren
Uncovering	#ontdekt
Victims	#patiënten
virtual experience	#virtuele wereld
Volunteers	#vrijwilligers

4a. The Hunt for the Higgs – TExSIS glossary

Anything	bestaan
Atlas	Atlas
Atoms	atomen
Began	begon
Big Bang	oerknal
Blip	afwijking
Born	geboren
CMS	CMS
completed Standard Model	Standaardmodel van elementaire deeltjes
Created	higgsdeeltje
Created	ontstaan
Deep	diepgaande
deep paradox at work	diepgaande paradox
Distinction	verschil
distinction between gravity	verschil tussen zwaartekracht
Elementary	elementaire
Encouragement	superdeeltjes
Energy	zone
Equations	vergelijkingen
Error	hoeveelheden
Existence	bestaan
Field	BEH-veld
Fluctuation	schommeling
Fluke	meetfouten
Force	krachten
Forces	krachten
Forces	natuurkrachten
Fraction	fractie
fraction of a second	fractie van een seconde
fundamental forces of nature	fundamentele natuurkrachten
GEV	giga-elektronvolt
Gravity	zwaartekracht
Groups	groepen
hunt for the Higgs	jacht op het BEH-deeltje
Idea	idee

idea of symmetry	idee van symmetrie
Incredible	gigantische hoeveelheid energie
LHC	LHC
Light	licht
Mass	massa
mass of the Higgs	massa van het higgsdeeltje
Masses	massa
Mathematics	wiskunde
Matter	materie
Nuclei	atoomkernen
Particles	deeltjes
Perfection	perfectie
Physics	natuurkunde
Pieces	onderdelen
Pioneers	pioniers
pioneers of a powerful new mathematical theory	pioniers van een machtige nieuwe wiskundige theorie
Predecessor	voorganger
Research	Research
Resolved	vinden
Results	resultaten
Richard	verantwoordelijk
Second	seconde
Sense	boson
Small	minieme
Sparticles	spartikels
Squarks	squarks
Standard Model	Standaardmodel
Standard Model of Elementary Particles	Standaardmodel voor elementaire deeltjes
Super Symmetry	supersymmetrie
Symmetry	symmetrie
Theme	kenmerk
Theorists	theoretici
Theory	theorie
Tiny	kleine
tiny fluctuation	kleine schommeling
Universe	heelal
Vindication	Standaardmodel voor elementaire deeltjes
Weak	zwakke
Works	werkt
Works	functioneert

4b. The Hunt for the Higgs – Gold Standard glossary

antimatter	deeltjes
Area	energiezone
Atlas detector	Atlas

Atlas detector	Atlasdetector
Atlas	Atlas
Atoms	atomen
Atoms	atomen
big bang	oerknal
Boson	deeltje
Breakage	verbreking
building blocks	basisbouwstenen
CERN	CERN
clumping together	samenklonteren
CMS	CMS
CMS	CMS
Collides	botsen
colliding particles	deeltjes * botsen
Colliding	botsen
Collisions	deeltjesbotsingen
Competition	concurrentie
Condensed	condenseert
Cosmos	heelal
create * disturbance	verstoren
Created	ontstaan
Created	verschijnt
Creation	oerknal
Creation	ontstaan
data plots	onderzoek
Data	gegevens
Data	meetgegevens
Description	theorie
Detected	opsporen
Detector	detector LHC-B
detectors * CERN	CERN-detectoren
DNA	DNA
Earth	aarde
ebbed out	rustiger
electromagnetic force	elektromagnetische kracht
Electromagnetic	elektromagnetische
electro-magnetism	elektromagnetisme
elementary particles	elementaire deeltjes
Elementary	elementaire
Elementary	elementaire
energy levels	hoeveelheid energie
energy range	zone
energy window	energiezone
Energy	energie
Equations	vergelijkingen

European Organisation for Nuclear Research	Europese Raad voor Nucleaire Research
existence	bestaat
experiments	experiment
Field	BEH-veld
Field	energieveld
fluctuation	schommeling
force particle	krachtvoerend deeltje
force partner	krachtvoerende partner
Force	kernkracht
Force	kracht
Force	krachtvoerend
forces of nature	natuurkrachten
Forces	energie
Forces	krachten
Forces	natuurkrachten
Form	Vorm
fragments	fragmenten
fundamental laws	basisnatuurwetten
Geneva	Genève
GEV energy window	energiezone * giga-elektronvolt
GEV	GeV
GEV	giga-elektronvolt
giga electron volts	giga-elektronvolt
Gluons	gluonen
God Particle	godsdeeltje
Gravity	zwaartekracht
Gravity	zwaartekracht
heat and fury	toestand
Here	Genève
Higgs Boson	Brout-Englert-Higgsdeeltje
Higgs Boson	deeltje
Higgs Boson	higgsboson
Higgs boson	higgsdeeltje
Higgs field	BEH- * higgsveld
Higgs field	BEH- *veld
Higgs field	BEH-veld
Higgs field	higgsveld * BEH-veld
Higgs field	higgsveld
Higgs hunters	bosonjagers
Higgs hunters	Brout, Englert * Higgs
Higgs hunters	onderzoekers
Higgs	BEH- * higgsboson
Higgs	BEH-boson
Higgs	BEH-deeltje
Higgs	boson

Higgs	Brout-Englert-Higgsdeeltje
Higgs	deeltje
Higgs	higgs- * BEH-boson
Higgs	higgsboson
Higgs	higgsbosonjagers
Higgs	higgsdeeltje
Higgs	higgsveld
History	geschiedenisboeken
Hunt	jacht
Big Bang	oerknal
It	CERN
It	higgsdeeltje
Large Hadron Collider	deeltjesversneller
Large Hadron Collider's	deeltjesversneller
laws of nature	natuurwetten
LEP Collider	LEP-versneller
Levels	hoeveelheid
LHC	deeltjesversneller
LHC	LHC
LHC's * detectors	detectoren
lower limit	benedengrens
Mass	massa
Masses	massa
mathematical theory	wiskundige theorie
Mathematical	wiskundige
Mathematics	wiskunde
matter particle	materiedeeltje
matter particles	materiedeeltjes
matter partner	materiepartner
Matter	materie
Measurements	metingen
MIT	MIT
Model	Standaardmodel
Model	theorie
Moment	seconde
natural world	natuur
nerve centre	zenuwcentrum
Nobel Prize	Nobelprijs
Noise	ruis
nuclei of atoms	atoomkernen
One	detector
particle accelerator	deeltjesversneller
particle accelerators	deeltjesversnellers
Particle	deeltje
Particles	deeltjes
Partner	partner

Perfection	perfectie
photinos	fotino’s
Photon	fotonen
physicists	natuurkundigen
physicists	Ze
Physics	natuurkunde
Piece	deeltje
Prizes	prijzen
proof of the existence	bevestigt
Protons	protonen
question	inzicht
radio-activity	radioactiviteit
Results	resultaten
scientists	Higgs * Belgen Brout en Englert
Scientists	mensen
scientists	Natuurkundigen
scientists	onderzoekers
scientists	wetenschappers
scientists’ equations	wetenschappelijke vergelijkingen
scientists’	wetenschappelijke
Search	zoekgebied
Second	momenten
Second	seconde
signature	energiepatroon
slowed down	vertraagd
sparticles	spartikels
sparticles	superdeeltjes * spartikels
sparticles	superdeeltjes
speed of light	lichtsnelheid
speed of light	snelheid van het licht
Squarks	squarks
Standard Model of Elementary Particles	Standaardmodel voor elementaire deeltjes
Standard Model	Standaardmodel van elementaire deeltjes
Standard Model	Standaardmodel
Stars	sterren
statistical	statistische
substance	substantie
Super Symmetric particles	supersymmetrische deeltjes
super symmetric twins	supersymmetrische tweelingen
super symmetric	supersymmetrische
Super Symmetry	supersymmetrici
Super Symmetry	supersymmetrie
symmetric	symmetrisch
symmetry	symmetrie
Test	uitgetest
Them	superdeeltjes

Them	supersymmetrische voorspellingen
theoretical physicists	theoretici
theoretical physicists	theoretisch natuurkundigen
theoretical physics	theoreticus
Theoretical	theoretisch
Theorists	theoretici
Theorists	Wetenschappers
Theory	theorie
Theory	theorieën
Twins	tweelingen
Underground	onder de grond
Universe	heelal
Universe	Standaardmodel voor elementaire deeltjes
Universe	universum
University College London	University College * Londen
upper limit	bovengrens
very beginning	ontstaan
Viruses	virussen
W * Z bosons	W-bosonen * Z-bosonen
w bosons	W-bosonen
z bosons	Z-bosonen

Appendix 3

The source texts for the experiments

1a. The Earth Machine – TExSIS group

THE EARTH MACHINE ***(over de geologische vorming van de aarde)***
A new type of lava erupted and burst through cracks in the crust.

Beneath the streets of the sunshine state *(Florida)* lies the Floridian Aquifer.

A precious resource that provides most of Florida's fresh drinking water. Up to 3 billion gallons a day.

However, when two tectonic plates crunch together.....they create huge splits in the ground called faults.

To make the stadium *(Memorial Stadium, home to the Golden Bears)* earthquake proof the Bears, will put two sections of the stadium on separate free-floating blocks of concrete so that if a quake hits, the stadium will survive.

The granite rock found here is called igneous.

Then there are sedimentary rocks like limestone.

The third kind of rock is called metamorphic.

If we remove a piece of the crust we can see the furnace that produces it. This is the mantle.

An enormous upwelling of intense heat –a mantle plume – is rising from the interior of the planet under this part of the Rift Valley.

And as soon as Ken changes the magnetic field – the turtle changes direction. So the turtles really are sensitive to the magnetic field. It's down to a mineral called magnetite.

This is the final destination. The centre of the earth.

It is called the inner core. It too is made of metal but this is now solid, crushed by pressures around four million times greater than on the surface of the planet.

1b. The Earth Machine – Gold Standard group

THE EARTH MACHINE *(over de geologische vorming van de aarde)*
A new type of lava erupted and burst through cracks in the crust.

Beneath the streets of the sunshine state *(Florida)* lies the Floridian Aquifer.

A precious resource that provides most of Florida's fresh drinking water. Up to 3 billion gallons a day.

However, when two tectonic plates crunch together.....they create huge splits in the ground called faults.

To make the stadium *(Memorial Stadium, home to the Golden Bears)* earthquake proof the Bears, will put two sections of the stadium on separate free-floating blocks of concrete so that if a quake hits, the stadium will survive.

The granite rock found here is called igneous.

Then there are sedimentary rocks like limestone.

The third kind of rock is called metamorphic.

If we remove a piece of the crust we can see the furnace that produces it. This is the mantle.

An enormous upwelling of intense heat –a mantle plume – is rising from the interior of the planet under this part of the Rift Valley.

And as soon as Ken changes the magnetic field – the turtle changes direction. So the turtles really are sensitive to the magnetic field. It's down to a mineral called magnetite.

This is the final destination. The centre of the earth.

It is called the inner core. It too is made of metal but this is now solid, crushed by pressures around four million times greater than on the surface of the planet.

2a. Madagascar, The Island of Marvels – TExSIS group

MADAGASCAR *(over de fauna en flora op Madagascar)*
They are lemurs. There are 80 different types, from nocturnal, mouse-sized creatures to this, the biggest, the size of a child. It's an indri.

As well as being Madagascar's equivalent of hedgehogs, tenrecs also take the place that moles and shrews would occupy anywhere else in the world.

It's a giraffe necked weevil, and this is a male.

A male panther chameleon – one of the biggest.

A pygmy chameleon, the world's tiniest reptile, tiptoes through the leaf litter on the steep volcanic slopes.

This is the Lac Alaotra reed lemur.

The seabed was pushed up, creating a great block of limestone. Over time, it's been carved by water into forests of giant pinnacles. This is the tsingy - one of Madagascar's strangest landscapes.

But crowned lemurs are as good at rock climbing as they are at tree climbing.

The fossa! No big African predators made it to Madagascar.

Instead the island's top predator is a giant mongoose.

A sunbird has become a nectar thief.

These are sifakas.

The drongo isn't even stealing the material – just chasing the flycatchers from their territory.

A rare Verreaux's coua, found only round this lake, puffs itself up until it's almost spherical.

It's Grandidier's vontsira, one of the world's rarest carnivores.

And on these windswept cliffs there are radiated tortoises, one of the world's most beautiful species.

2b. Madagascar, The Island of Marvels – Gold Standard group

MADAGASCAR *(over de fauna en flora op Madagascar)*
They are lemurs. There are 80 different types, from nocturnal, mouse-sized creatures to this, the biggest, the size of a child. It's an indri.

As well as being Madagascar's equivalent of hedgehogs, tenrecs also take the place that moles and shrews would occupy anywhere else in the world.

It's a giraffe necked weevil, and this is a male.

A male panther chameleon – one of the biggest.

A pygmy chameleon, the world's tiniest reptile, tiptoes through the leaf litter on the steep volcanic slopes.

This is the Lac Alaotra reed lemur.

The seabed was pushed up, creating a great block of limestone. Over time, it's been carved by water into forests of giant pinnacles. This is the tsingy - one of Madagascar's strangest landscapes.

But crowned lemurs are as good at rock climbing as they are at tree climbing.

The fossa! No big African predators made it to Madagascar.

Instead the island's top predator is a giant mongoose.

A sunbird has become a nectar thief.

These are sifakas.

The drongo isn't even stealing the material – just chasing the flycatchers from their territory.

A rare Verreaux's coua, found only round this lake, puffs itself up until it's almost spherical.

It's Grandidier's vontsira, one of the world's rarest carnivores.

And on these windswept cliffs there are radiated tortoises, one of the world's most beautiful species.

3a. The Secret World of Pain – TExSIS group

THE SECRET WORLD OF PAIN *(over oorzaken en gevolgen van (chronische) pijn)*
John believes the Marsillis *(a family of people who don't feel pain)* have a faulty gene which somehow affects how their sensory neurons grow and develop.

If these neurons are defective, the pathways along which pain signals flow don't work properly.

They *(scientists)* analysed their *(the Marsillis)* DNA and compared it with the general population to look for differences and discovered it was caused by mutations in one single gene: the SCN9a.

It regulates the electrical signals which transmit the sensation of pain to the brain.

Rachel suffers from Complex Regional Pain Syndrome.

For Maria it is these repeated painful treatments that could be altering the normal development of their pain pathways.

These will ultimately shape how our pain system is wired up.

Irene first showed her unsuspecting volunteers a triangle and at the same time gave them a low pain stimulus.

Areas of Jonathan's brain linked to decision-making, emotion and attention transmitted messages to the brain stem.

This area of the brain acts as a gatekeeper, blocking the pain signals from entering his brain.

By manipulating that part of the brain's attentional system, Khalib will enter the virtual experience of Snow World, a world far away from his pain and suffering.

Every pain experience we have involves many different parts of the brain creating its own neural signature.

And it is this pain, pain that persists long after the injury is healed and serves no purpose at all, that is known as chronic pain.

Rebecca has come to the Pain Relief Foundation for trans-cranial magnetic stimulation, a treatment for chronic pain which is still in clinical trials.

Turo's theory is that the constant barrage of pain signals experienced by Rebecca could be causing changes in the part of her brain that is responsible for movement, the motor cortex.

By giving her repeated magnetic pulses, Turo believes this should help re-wire the nerve cells back to their normal function, correcting her motor cortex which should help reduce her pain.

3b. The Secret World of Pain – Gold Standard group

THE SECRET WORLD OF PAIN *(over oorzaken en gevolgen van (chronische) pijn)*
John believes the Marsillis *(a family of people who don't feel pain)* have a faulty gene which somehow affects how their sensory neurons grow and develop.

If these neurons are defective, the pathways along which pain signals flow don't work properly.

They *(scientists)* analysed their *(the Marsillis)* DNA and compared it with the general population to look for differences and discovered it was caused by mutations in one single gene: the SCN9a.

It regulates the electrical signals which transmit the sensation of pain to the brain.

Rachel suffers from Complex Regional Pain Syndrome.

For Maria it is these repeated painful treatments that could be altering the normal development of their pain pathways.

These will ultimately shape how our pain system is wired up.

Irene first showed her unsuspecting volunteers a triangle and at the same time gave them a low pain stimulus.

Areas of Jonathan's brain linked to decision-making, emotion and attention transmitted messages to the brain stem.

This area of the brain acts as a gatekeeper, blocking the pain signals from entering his brain.

By manipulating that part of the brain's attentional system, Khalib will enter the virtual experience of Snow World, a world far away from his pain and suffering.

Every pain experience we have involves many different parts of the brain creating its own neural signature.

And it is this pain, pain that persists long after the injury is healed and serves no purpose at all, that is known as chronic pain.

Rebecca has come to the Pain Relief Foundation for trans-cranial magnetic stimulation, a treatment for chronic pain which is still in clinical trials.

Turo's theory is that the constant barrage of pain signals experienced by Rebecca could be causing changes in the part of her brain that is responsible for movement, the motor cortex.

By giving her repeated magnetic pulses, Turo believes this should help re-wire the nerve cells back to their normal function, correcting her motor cortex which should help reduce her pain.

4a. The Hunt fort he Higgs – TExSIS group

THE HUNT FOR THE HIGGS

Just before Christmas researchers working at CERN near Geneva announced that they'd caught a tantalising glimpse of the Higgs Boson. The search for the Higgs takes us deep into the most important questions about how the universe works and how it was created. The problem with hunting for the Higgs is that it can't be detected in everyday conditions. To find it, scientists need to return to those at the very beginning. Well almost, to the conditions just after the big bang when the theory goes, the Higgs and everything else was first created.

In this one crucial second all the elementary particles were created including, scientists believe, the Higgs Boson.

The Large Hadron Collider's technique to transport scientists to the moment just after the Big Bang is as violent as it is ambitious.

A hundred metres underground it takes protons from the nuclei of atoms and collides them at almost the speed of light.

It would be proof of the existence of a field that scientists believe surrounds us all the time and that appeared in that first second of creation. As the heat and fury ebbed out of the Big Bang so the theory goes, the Higgs field condensed. As particles travelled through this field they get slowed down like travelling through treacle, this is what gives them mass. Without gaining mass, particles would have continued to fly through the universe at the speed of light, never clumping together to form you, me, blackboards – well, anything. [...] Scientists need to create a disturbance in the Higgs field to detect the boson itself. This is what the LHC is attempting to do, by colliding particles. It's a challenge other particle accelerators have tried and been unable to complete because for all scientists sense that the Higgs ought to be there, it has proven too spectacularly difficult to find. What's made all the difference at the LHC are the incredible energy levels the collider can reach, pushing further back in time into that crucial first second.

This has opened up new places to search for the Higgs, a hunt that's defined in terms of what mass the Higgs itself might have, measured in GEV or giga electron volts.

After decades of work, the LEP Collider at CERN - a predecessor of the LHC - ruled out the Higgs being at the bottom end of potential masses. And by November 2011the LHC had already radically narrowed the search. In November, that left a region of just 30 GEV for the Higgs to be hiding in. But this last remaining energy range is also the trickiest to search.

It's the area in which the unique signature of the Higgs is most deeply buried under the background noise of other particles created in the collider.

The experimental physicists here at CERN have already put some of the ideas of their colleagues, the theorists, to the test and not all the results have been positive. But what's at stake with the Higgs isn't just one particle, however elusive, or any old theory. The Higgs is the cornerstone for the most successful and all-encompassing description of how our universe works that there is. Working this beautiful model out has been one of the great achievements of theoretical physics, and Frank Wilczek was one of the key contributors. And all that puzzling won Frank a Nobel Prize for his contribution to what's called the Standard Model of Elementary Particles. The Standard Model is essentially an understanding of how all the pieces of the universe fit together except for gravity, a mind boggling project.

These particles are more like lumps of energy and they transmit the forces that bring the matter particles to life, like the photon which carries the electromagnetic force.

The gluons that carry the strong force which holds the nuclei of atoms together and W and Z bosons that are responsible for the weak force governing radio-activity.

[...]

Scientists plan to use the completed Standard Model as the foundation for an even deeper description of the universe, one based on the idea of symmetry and its breakage.

James became one of the pioneers of a powerful new mathematical theory called Super Symmetry. Using symmetry in equations had previously led to the discovery of antimatter. These new ones suggested there was another hidden world of particles no one had suspected. The theory took everything we thought we knew about even the Higgs and doubled it, giving every matter particle a force partner, and every force particle a matter partner. These heavier super symmetric twins were labelled sparticles. According to Super Symmetry, matter and forces aren't so distinct after all. There's a grand symmetry between them that we can currently see only one partner from each pair. However strange it seems, this theory has gained widespread support from theoretical physicists. Not just for the beauty of its equations but for what it might help explain. That's where the six billion pound experiments at CERN may really usher in a revolution. ~~~ Because they're hunting for evidence of Super Symmetry. Richard Jacobsson is in charge of the operation of the detector that may give the first clues about Super Symmetric particles.

So far, not only have they found no evidence of the photinos, squarks or other sparticles predicted by the theorists, they've even ruled out the possibility of them that some of the energies theorists were hoping they'd be.

4b. The Hunt fort he Higgs – Gold Standard group

THE HUNT FOR THE HIGGS *(about the research for the Higgs boson).*

Just before Christmas researchers working at CERN near Geneva announced that they'd caught a tantalising glimpse of the Higgs Boson. The search for the Higgs takes us deep into the most important questions about how the universe works and how it was created. The problem with hunting for the Higgs is that it can't be detected in everyday conditions. To find it, scientists need to return to those at the very beginning. Well almost, to the conditions just after the big bang when the theory goes, the Higgs and everything else was first created.

In this one crucial second all the elementary particles were created including, scientists believe, the Higgs Boson.

The Large Hadron Collider's technique to transport scientists to the moment just after the Big Bang is as violent as it is ambitious.

A hundred metres underground it takes protons from the nuclei of atoms and collides them at almost the speed of light.

It would be proof of the existence of a field that scientists believe surrounds us all the time and that appeared in that first second of creation. As the heat and fury ebbed out of the Big Bang so the theory goes, the Higgs field condensed. As particles travelled through this field they get slowed down like travelling through treacle, this is what gives them mass. Without gaining mass, particles would have continued to fly through the universe at the speed of light, never clumping together to form you, me, blackboards – well, anything. […] Scientists need to create a disturbance in the Higgs field to detect the boson itself. This is what the LHC is attempting to do, by colliding particles. It's a challenge other particle accelerators have tried and been unable to complete because for all scientists sense that the Higgs ought to be there, it has proven too spectacularly difficult to find. What's made all the difference at the LHC are the incredible energy levels the collider can reach, pushing further back in time into that crucial first second.

This has opened up new places to search for the Higgs, a hunt that's defined in terms of what mass the Higgs itself might have, measured in GEV or giga electron volts.

After decades of work, the LEP Collider at CERN - a predecessor of the LHC - ruled out the Higgs being at the bottom end of potential masses. And by November 2011the LHC had already radically narrowed the search. In November, that left a region of just 30 GEV for the Higgs to be hiding in. But this last remaining energy range is also the trickiest to search.

It's the area in which the unique signature of the Higgs is most deeply buried under the background noise of other particles created in the collider.

The experimental physicists here at CERN have already put some of the ideas of their colleagues, the theorists, to the test and not all the results have been positive. But what's at stake with the Higgs isn't just one particle, however elusive, or any old theory. The Higgs is the cornerstone for the most successful and all-encompassing description of how our universe works that there is. Working this beautiful model out has been one of the great achievements of theoretical physics, and Frank Wilczek was one of the key contributors. And all that puzzling won Frank a Nobel Prize for his contribution to what's called the Standard Model of Elementary Particles. The Standard Model is essentially an understanding of how all the pieces of the universe fit together except for gravity, a mind boggling project.

These particles are more like lumps of energy and they transmit the forces that bring the matter particles to life, like the photon which carries the electromagnetic force.

The gluons that carry the strong force which holds the nuclei of atoms together and W and Z bosons that are responsible for the weak force governing radio-activity.

[…]

Scientists plan to use the completed Standard Model as the foundation for an even deeper description of the universe, one based on the idea of symmetry and its breakage.

James became one of the pioneers of a powerful new mathematical theory called Super Symmetry. Using symmetry in equations had previously led to the discovery of antimatter. These new ones suggested there was another hidden world of particles no one had suspected. The theory took everything we thought we knew about even the Higgs and doubled it, giving every matter particle a force partner, and every force particle a matter partner. These heavier super symmetric twins were labelled sparticles. According to Super Symmetry, matter and forces aren't so distinct after all. There's a grand symmetry between them that we can currently see only one partner from each pair. However strange it seems, this theory has gained widespread support from theoretical physicists. Not just for the beauty of its equations but for what it might help explain. That's where the six billion pound experiments at CERN may really usher in a revolution. Because they're hunting for evidence of Super Symmetry. Richard Jacobsson is in charge of the operation of the detector that may give the first clues about Super Symmetric particles.

So far, not only have they found no evidence of the photinos, squarks or other sparticles predicted by the theorists, they've even ruled out the possibility of them that some of the energies theorists were hoping they'd be.

Appendix 4

Instructies voor het experiment zonder terminologielijsten[109]

INSTRUCTIES EXPERIMENT GSM AFZETTEN AUB[110]

- Je bent ingelogd met een guestaccount[111]

- Open Google en de Van Dales E-Nl / Nl-Nl indien nog niet gebeurd vooraleer je begint.

- Bronteksten (in .doc) staan op bureaublad

- Instellingen in Inputlog invullen:

'Session identification'
Participant: je naam
Text Language: NL
Age: je leeftijd
Gender: 1=m 2=vr
Session: 1B= brontekst 1 / 2B= brontekst 2 / 3B=brontekst 3
Group/Experience: niets invullen

- Vertaling:

- titel niet vertalen. Wat cursief tss haakjes staat ook niet, dient als inhoudelijke ondersteuning.
- **DADELIJK ALLE ZINNEN GOED VERTALEN IN DE VOLGORDE WAARIN ZE STAAN, GEEN REVISIE ACHTERAF.**
- gebruik alle onlinebronnen die je nodig acht.

---

[109] Translation in english below.

[110] Not mentioned for the professionals, it was asked orally whether they could refrain from answering calls during the experiment.

[111] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.

- pas overgaan naar tekst 2 als tekst 1 helemaal af is, idem tekst 3.
- doseer de tijd, je krijgt 2u30, probeer de drie teksten volledig te vertalen. Als dat niet lukt, werk dan niet haastig naar het einde toe, blijf goed vertalen en tijd nemen voor opzoekingswerk (students' version). // er is geen tijdsbeperking maar het is wel de bedoeling dat je de 4 tekstjes na elkaar vertaalt (professionals' version).

- Starten: klik op 'record' - er opent zich een worddocument **(.WORDLOG), DAARIN SCHRIJF JE JE VERTALING.**

- Klaar met tekst 1: klik 1x op inputlog-icoontje in de balk onderaan (indien geen reactie klik dan nog eens); dan op 'STOP RECORDING', **bij 'UNSAVED CHANGES': klik NO.**
Mij verwittigen zodat ik de instellingen kan aanpassen (of zelf doen). Idem na tekst 2.

- **Volledig klaar:** pc niet afsluiten, ik kom de bestanden downloaden en ik sluit zelf je pc af.[112]

(translation)

Instructions for the experiment without glossaries

INSTRUCTIONS EXPERIMENT PLEASE TURN OFF YOUR MOBILE PHONE[113]

- You are logged with a guestaccount[114]

- Open Google and the Van Dale dictionaries E-Dutch / Dutch – Dutch if not already done before you start.

- Source texts (in .doc) are on your desktop

- Please fill in the settings in Inputlog:

'Session identification'
Participant: your name

[112] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.
[113] Not mentioned for the professionals, it was asked orally whether they could refrain from answering calls during the experiment.
[114] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.

Text Language: Dutch
Age: your age
Gender: 1=m 2=f
Session: 1B= source text 1 / 2B= source text 2 / 3B=source text 3
Group/Experience: not to be filled in

- Translation:

- don't translate the title. Italics between brackets neither, this is extra content information to better understand the sentence.
- **PLEASE MAKE A GOOD TRANSLATION AT ONCE WITHOUT REVISION AT THE END OF THE TASK AND RESPECT THE ORDER IN WHICH THE SENTENCES ARE PRESENTED.**
- use all the online sources you need.
- start with text 2 only when you have finished text 1 and with text 3 when text 2 is finished.
- balance your time, you have 2h30, try to translate all three the texts but if you don't make it, don't rush, keep on making a good translation and take your time for all the research you need to do (students' version). // there is no set time limit but you need to translate all four the texts (professionals' version).

- To start: click 'record' – a word document is opened **(.WORDLOG), THIS IS THE DOCUMENT WHERE YOU NEED TO WRITE DOWN YOUR TRANSLATION.**

- When you are ready with text 1: click once the inputlog icon on the bar below (if no reaction click again); then click 'STOP RECORDING', **if 'UNSAVED CHANGES': click NO.**
Call me so I can adapt the settings (or do it your self). Idem after text 2 and 3.

When you are finished: don't shut down the pc, I have to download the files. I will shut down the pc[115].

[115] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.

Instructies voor het experiment met terminologielijsten[116]

INSTRUCTIES EXPERIMENT GSM AFZETTEN AUB[117]

- Je bent ingelogd met een guestaccount[118]

- Open Google en de Van Dales E-Nl / Nl-Nl indien nog niet gebeurd vooraleer je begint.

- Bronteksten (in .doc) + terminologielijsten (in .xlsx) staan op bureaublad

- Instellingen in Inputlog invullen:

'Session identification'
Participant: je naam
Text Language: NL
Age: je leeftijd
Gender: 1=m 2=vr
Session: 1B= brontekst 1 / 2B= brontekst 2 / 3B=brontekst 3
Group/Experience: niets invullen

- Vertaling:

- titel niet vertalen. Wat cursief tss haakjes staat ook niet, dient als inhoudelijke ondersteuning.
- **DADELIJK ALLE ZINNEN GOED VERTALEN IN DE VOLGORDE WAARIN ZE STAAN, GEEN REVISIE ACHTERAF.**
- gebruik alle onlinebronnen die je nodig acht, de terminologielijsten staan op je bureaublad.

*De **terminologielijsten** (alfabetisch gerangschikt; in een excelbestand) geven NIET de Engelse term met de Nederlandse woordenboekvertaling maar de overeenkomstige term uit de bestaande, officiële, correcte vertaling.*
De term uit de lijst is een optie, een suggestie, je moet zelf uitmaken of je die wilt gebruiken of niet.
De termen zijn onderlijnd in de brontekst zoals je ze vindt in de lijsten: meerdere woorden samen onderlijnd (met een ononderbroken lijn) = vind je samen terug in de lijst.
***In de terminologielijst zoeken met CRTL+F** zodat Inputlog registreert welke term je zocht.*

- pas overgaan naar tekst 2 als tekst 1 helemaal af is, idem tekst 3.

[116] Translation in english below.
[117] Not mentioned for the professionals, it was asked orally whether they could refrain from answering calls during the experiment.
[118] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.

- doseer de tijd, je krijgt 2u30, probeer de drie teksten volledig te vertalen. Als dat niet lukt, werk dan niet haastig naar het einde toe, blijf goed vertalen en tijd nemen voor opzoekingswerk.

- Starten: klik op 'record' - er opent zich een worddocument **(.WORDLOG), DAARIN SCHRIJF JE JE VERTALING.**

- Klaar met tekst 1: klik 1x op inputlog-icoontje in de balk onderaan (indien geen reactie klik dan nog eens); dan op 'STOP RECORDING', **bij 'UNSAVED CHANGES': klik NO.**
Mij verwittigen zodat ik de instellingen kan aanpassen (of zelf doen). Idem na tekst 2.

- **Volledig klaar:** pc niet afsluiten, ik kom de bestanden downloaden en ik sluit zelf je pc af[119].

(translation)

Instructions for the experiment with glossaries

INSTRUCTIONS EXPERIMENT PLEASE TURN OFF YOUR MOBILE PHONE[120]

- You are logged with a guestaccount[121]

- Open Google and the Van Dale dictionaries E-Dutch / Dutch – Dutch if not already done before you start.

- Source texts (in .doc) and glossaries (in .xlsx) are on your desktop

- Please fill in the settings in Inputlog:

'Session identification'
Participant: your name
Text Language: Dutch
Age: your age
Gender: 1=m 2=f
Session: 1B= source text 1 / 2B= source text 2 / 3B=source text 3
Group/Experience: not to be filled in

[119] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.
[120] Not mentioned for the professionals, it was asked orally whether they could refrain from answering calls during the experiment.
[121] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.

- Translation:

- don't translate the title. Italics between brackets neither, this is extra content information to better understand the sentence.
- PLEASE MAKE A GOOD TRANSLATION AT ONCE WITHOUT REVISION AT THE END OF THE TASK AND RESPECT THE ORDER IN WHICH THE SENTENCES ARE PRESENTED.
- use all the online sources you need, the glossaries are on your desktop.

*The **glossaries** (in alphabetic order; in an excelsheet) DON'T provide the English term with a Dutch dictionary translation but the corresponding term used in the existing, official, correct translation.*
The term from the glossary is an option, a suggestion, you have to decide yourself whether to use it or not.
The terms are underlined in the source tekst as they are represented in the glossaries: more then one word underlined (with an uninterrupted line) = you can find these words together in the glossary.
***Use CRTL+F to search for terms in the glossaries,** doing so Inputlog registers which term you were looking for.*

- start with text 2 only when you have finished text 1 and with text 3 when text 2 is finished.
- balance your time, you have 2h30, try to translate all three the texts but if you don't make it, don't rush, keep on making a good translation and take your time for all the research you need to do (students' version). // there is no set time limit but you need to translate all four the texts (professionals' version).

- To start: click 'record' – a word document is opened (.WORDLOG), THIS IS THE DOCUMENT WHERE YOU NEED TO WRITE DOWN YOUR TRANSLATION.

- When you are ready with text 1: click once the inputlog icon on the bar below (if no reaction click again); then click 'STOP RECORDING', if 'UNSAVED CHANGES': click NO.
Call me so I can adapt the settings (or do it your self). Idem after text 2 and 3.
When you are finished: don't shut down the pc, I have to download the files. I will shut down the pc.[122]

[122] Not mentioned for the professionals as the experiment was conducted at their own working place with a university laptop.