**DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE**

COMPUTER-ASSISTED TRANSLATION AT THE SERVICE OF TRANSLATORS

A Unified Methodology

Daniel Vidal Hussey

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2015
Supervisor:	Aarne Ranta
Examiner:	Robin Cooper
Report number:	(number will be provided by the administrators)
Keywords:	computer-assisted translation, grammatical framework, style guide, post-editing

COMPUTER-ASSISTED TRANSLATION AT THE SERVICE OF TRANSLATORS

A Unified Methodology

DANIEL VIDAL HUSSEY

Master's Programme in Language Technology
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg

September 2015

SUPERVISOR:
Aarne Ranta

EXAMINER:
Robin Cooper

No wise fish would go anywhere without a porpoise.

— Alice in Wonderland

ABSTRACT

This thesis identifies shortcomings with how Computer-Assisted Translation (CAT) tools are developed. Their final use as an aid to translators is often not fully considered and left to others to evaluate. A unified methodology is proposed which allows a CAT tool to be evaluated intrinsically and extrinsically using methods that show the tool's effect on the whole translation process. Special emphasis is placed on prototyping as a resource-effective way to create tools and gain critical feedback before a full implementation. To evaluate the methodology's usefulness, StyleCheck is developed and evaluated using it. StyleCheck is a tool implemented in Grammatical Framework. It detects when a style guide rule is applied and gives a hint to the translator when it isn't. The use of methods developed in Translation Process Research generates a wealth of data that gives detailed insights into how a tool performs. Results show StyleCheck is effective at getting a style guide to be applied, more than translating from scratch or post-editing, although more work on the user interface is required. The methodology is proven to be good at coming up with CAT tool improvements, quickly protoyping them and evaluating them.

ACKNOWLEDGEMENTS

Many people deserve a big "Thank you!" for their help in one way or another. I'll mention a few so as not to keep readers too long from the rest of the thesis.

Thank you:

- First, to my supervisor Aarne Ranta for helpful and patient insight.
- Second, to all the participants in the case study who took the time to thoughtfully fill in the questionnaires and translate the texts.
- Last and perhaps most importantly, to all the people who asked what my thesis was about and gave me puzzled looks when I told them. After explaining it to each of them in a different way, hopefully the following pages will have found the ideal answer.

CONTENTS

ACRONYMS

CAT Computer-Assisted Translation

CNL Controlled Natural Language

GF Grammatical Framework

HT Human Translation

MT Machine Translation

NLP Natural Language Processing

PE Post-Editing

RGL Resource Grammar Library

SG Style Guide

SMT Statistical Machine Translation

ST Source Text

TM Translation Memory

TPR Translation Process Research

TS Translation Studies

TT Target Text

Part I

INTRODUCTION AND BACKGROUND

1

INTRODUCTION

Translation is the shared object of study of two research communities: Translation Studies (TS) and Machine Translation (MT), which is part of the larger Natural Language Processing (NLP) community. Despite this common interest, it comes as both a surprise and a concern that there is little to no interaction between these communities and their research. Each one approaches translation from a different perspective, using different methodological tools and with different aims. The most evident difference is on the definition of the object of study itself: translation. While in TS many a pages and thoughts have been put into trying to define what translation is and how both translators as subjects and translating as a process fit together to create the final product, in MT all this is taken for granted and nearly no thought is spared as to what it means to translate.

Computer-Assisted Translation (CAT), where translators use computers and specialised software to help them translate, seems a natural meeting point for these two communities. Indeed, in the early days of MT research, CAT was proposed as the way to go in the seminal ALPAC report (Pierce & Carroll, 1966), again years later by Kay (1980) and instisted upon yet again by Kay (1997). Despite these influential papers, only recently has strong research into CAT started picking up. Up until now, the field of CAT has seen the introduction of some important technologies such as the use of Translation Memory (TM) and glossaries. Apart from these two technologies, commercial CAT tools have mainly focused on the project management and team coordination aspects of translating.

With the recent improvement in perceived quality of MT output, especially after the introduction of the Moses SMT system (Koehn et al., 2007), CAT now seems attractive to MT researchers. The focus, however, has been on Post-Editing (PE). In PE, the translator is charged with nothing but "cleaning up" a machine translation to bring it up to acceptable quality. The story goes that this approach is faster and sufficient for certain texts that don't need full human translation quality, but this is not always the case. Humans are treated as simple quality assurance mechanisms instead of as people with a certain set of skills that could be put to good use. The move to PE seems to serve more the interests of claiming MT is viable and economical than the interests of improving translator's working conditions.

I argue that the problems with PE and with TM before it — and more generally of translation carried out between humans and computers — arise from the compartmentalised approach to CAT tool research

and development. NLP researchers approach the problem from their knowledge and skills, often forgetting that the systems they develop have to work well with people, a very different problem to algorithm development. Researchers can end up creating tools that not only interact poorly with translators, but also potentially work against them. Similarly, translators and researchers from TS are also weary of the claims that a machine can do what human translators can, to the point of taking over their jobs or a considerable portion of them. To this end, numerous studies (see Chapter 3) have tried to surface all the shortcomings of these approaches.

This situation is not desirable for either community. What is needed is to integrate both of them and collaborate in developing CAT tools. The basic step is to change development methodologies and move away from the current state where one community creates a tool and puts it out there, and the other has to evaluate it and pick holes in it.

This thesis proposes a research methodology (Chapter 4) for developing CAT tools with a tighter-knit integration of the skills of both communities. The gist is that theoretical research in TS and translator surveys should uncover needs which can potentially be addressed with NLP technologies. NLP experts can then develop a prototype technology which solves this problem. It does not necessarily have to be fully-fledged, but it should have the potential to be so. Finally, Translation Process Research (TPR) methodologies can be used to prove that this new technology actually affects translations and translators positively. If all goes well, a full implementation can be left to future research projects in academia or in commercial applications.

My hypothesis is that CAT tools developed on the basis of the proposed methodology should generate sufficient data to allow the tool to be evaluated instrinsically and extrinsically. In order to evaluate the methodology, in this thesis I use it to develop and evaluate StyleCheck, a tool that implements style guides. The results (Chapter 9) of the tool evaluation are presented, and a metaevaluation of the methodology is discussed.

It is hoped that through the use of this methodology, CAT tools developed in the future will be liked by translators and have a positive impact on the work they carry out. In a nutshell, CAT at the service of translators.

2

TRANSLATION, TRANSLATORS AND TRANSLATING

Before delving into CAT tools, it is best to start from the beginning and discuss what translation is. General theories of translation are estranged in MT research. Other areas of research in NLP base their research on theories — commonly linguistic theories —, upon which models are built, implemented and tested. MT and CAT systems today are not explicitely built on any general theory of translation.

In this chapter, some of the main theories of translation are discussed for two reasons. First and foremost, a strong theoretical basis will reinforce research into MT, CAT and related technologies as well as provide a better idea of where they should be heading and what can be improved. Having an explicit starting point for research can clear the way for competing theories to emerge and develop their own lines of research if considered necessary.

Second, Translation Process Research (TPR) research (Section 2.3) — part of theoretical research into translation — provides a useful set of methodological tools. These tools will be used in this thesis and integrated into the proposed research methodology (Chapter 4).

2.1 OVERVIEW OF PARADIGMS IN TRANSLATION THEORY

This section provides a high-level overview of the main paradigms in translation theory. They are taken as per Pym (2009), who describes five basic paradigms. Together they cover most of the individual theories of translation that have been proposed over the years.

2.1.1 *Theories of Equivalence*

The equivalence paradigm postulates that there is an equalness of value between the Source Text (ST) and the Target Text (TT) at some level. This paradigm relies heavily on linguistics, as can be seen in the work of Catford (1965), who postulated different possible equivalence levels for a ST and a TT: phonetic, lexical, syntactic, etc.

More generally, the focus in the theories of equivalence is placed on the ST, which guides and establishes what will appear in the TT. Translation is thus understood in the sense of creating a TT which is equivalent to the ST and contains nothing but the ST (save clearly delimited translator's notes). This is the notion that is the most widespread amongst the general public and MT research.

2.1.2 *Skopos or Functional Theories*

Skopos theory (from the Greek *skopos* 'aim', 'intention') introduced a radically new idea in translation: what matters when translating is the purpose, the function that the TT has to fulfil (Holz-Mänttäri & Tiedeakatemia, 1984; Reiß & Vermeer, 1984). This theory escaped from the constraints of linguistics and the ST to enter the real world, where translators carry out translations according to customers' requirements. Large variations in the TT are acceptable if justified by the function that the TT has to fulfil, which is often different from that of the ST. Equivalence is thus relagated to the case where the function is the same for the ST and the TT.

One of the most profound implications of this theory is that the reasons for choosing to translate a ST in one way or another cannot be found in linguistics, they are found in communication, ethics, sociology, marketing, etc. This lead to the appearance of agents in the paradigm, i. e. the customer, the employer, users of the translation, etc. For some theoreticians, the main agents in this paradigm are the translators, since they are the ones who have the final word on decisions that have to be made and other agents simply collaborate in this task (Reiß & Vermeer, 1984). For others, the focus is more on the client's requirements and not on the translator's individuality (Nord, 1997).

2.1.3 *Descriptive Theories*

Descriptive Translation Studies, developed in parallel to skopos theories, again presented a breakwaway from previous theories. In this paradigm the notion of "assumed translations" (Toury, 1995b) takes centre stage: all translations are equivalent due to them being considered a translation. Thus, the aim is to observe originals and translations and describe how they differ and what makes them equivalent rather than prescribing a certain type of equivalence or a need to adapt to the TT function as the only valid ways to translate.

One of the main notions in the descriptive paradigm is that of norms, whose main proponent is Toury (1995a). Translations are seen to be produced within a certain culture and society, and at a certain time in history. Norms turn the ideas and values within a society into "performance instructions", i. e. what can or cannot be done, what is permissable and what isn't (Toury, 1995a). Translators will thus adopt different strategies and produce different translations according to the society and historical time they live in. The notion of norms can lead to prescriptivisim: discovering the relevant norms allows for translations to be classified as good or bad depending on whether they follow the norms of a certain time and society. Thus,

the notion of norms can lead back to the presciptivism it was trying to escape.

2.1.4 *Theories of Indeterminacy*

The paradigm of indeterminism questions many of the assumptions of the other paradigms, and even the existance of translation itself. The principle of indeterminacy (Quine, 1969) postulated that it is impossible to know the real meaning of an utterance, there are always other possible interpretations based on the same empirical evidence. In translation, this led to the view that there are always many possible ways to translate a text and it is impossible to be certain that any one of them is right or is equivalent to the "meaning" of the ST. We can always construct hypotheses to defend one interpretation or another, all of which can be supported by the same empirical evidence, i. e. the ST (Quine, 1969).

Pym (2009) groups various related but somewhat different theories in this paradigm. A full overview is not within the scope of this thesis, but the notion of "abusive fidelity" (Lewis, 1985) is one worth mentioning. One of the problems with indeterminisim is that it doesn't provide translators with ways of carrying out their work. The closest there is to a guideline is to "abuse fidelity" and translate the key points within a text as close as possible to the ST, to the point of making the TT sound strange to the target culture (Lewis, 1985). Thus, users are made aware of the translation itself by breaking the illusion of equivalence or symmetry between languages (Snell-Hornby, 1988), the illusion that a universal transfer of meaning is possible.

2.1.5 *Localisation*

Pym (2009) introduces and describes localisation as a new paradigm in TS. The process of localisation — translating a product into different locales — introduces three important concepts. First, the idea of locale, which combines both a language and a specific culture. For example, Spanish can be divided into the Spain-Spanish locale, the Mexico-Spanish locale, etc. Second, internationalisation as the process of adapting a ST so that it can be easily translated into various locales. This entails the creation of an intermediate product that is either augmented (more space for strings, adding date formatting options to software, etc.) or simplified through the use of a Controlled Natural Language (CNL).

Third, Pym (2009) cites the rise of non-linear texts, which are often those that are localised. Non-linear texts are neither translated nor used from beginning to end. Rather, fragments are used such as the various strings in a piece of software, parts of a reference manual for software or appliances, etc. These products are usually incrementally

updated, meaning that future translations are simply the new or modified segments of the product. It often occurrs that these segments are presented to translators with little to no context as to where they are to appear in the final product.

The three previous factors bring about a change in the way translators work. Translation is but a single step in a larger process consisting of internationalisation, translation, editing and quality assurance. Translation is reduced to an artificial kind of equivalence weighed down by the importance of reuse and accepting Translation Memory suggestions or terminology from glossaries (even though translators may consider them wrong or inadequate), with little incentive for translators to want to improve it (Pym, 2009). Esselink (2000) notes that translators should carry out other steps in the localisation process, such as internationalisation and final editing, for which their intercultural skills are very useful and relevant.

2.2 MT: TRANSLATION AS A DISTRIBUTED SYSTEM

The paradigms presented in the previous section are but a brief glimpse of all the opinions and theories that have been proposed about translation, but a detailed review of them is outside of the scope and aim of this thesis[1]. What they do provide is a good basis for discussion of MT and how it fits in and relates to the paradigms.

Pym (2009) places MT in the equivalence paradigm, albeit a new kind of equivalence that serves the higher goal of reuse. Looking at the methods Statistical Machine Translation (SMT) uses today, this would seem to be the case. A large corpus of translated texts are collected, aligned at the sentence level, broken down into phrases or syntactic structures, aligned again and probabilities are calculated. Texts to be translated are then broken down into sentences and these pieces are then reused to create the TT. Just as was the case in the theories of equivalence, the ST rules the roost: it is the main element that guides what appears in the TT and linguistics is all that matters. Kay (1997) also shares this view of MT, as the following passage — which still rings true today — shows:

> Workers in this field [*MT research*] usually take it as self-evident what constitutes a translation, especially within the relatively limited domains in which automatic methods might be applied. In particular, it is assumed that the sequence of sentences that make up the translation should preserve the intended meanings of the corresponding sentences in the original.
> — Kay (1997)

1 For further details, Pym (2009) provides a thorough discussion of each paradigm, its authors, its pros and cons and its authors.

Čulo (2014) characterises MT as being instrumental and aiming to be functionally constant. Thus, MT is still stuck in the equivalence paradigm, but limited to cases where the ST and TT functions are the same. The problem arises when considering the corpora that SMT systems are based on: since they will have been translated by different people, from different sources and for different purposes, can we be certain that what resulting translations will be functionally identical? What is more, statistical systems will choose the most probable translation options from this amalgama, can probability really guarantee functional equivalence? From the point of view of descriptive theories, it might well be that the most probable translations that SMT generates correspond to translation norms present in the parallel corpus used. If, for example, a corpora of 19th century texts were fed to an MT system, the resulting translations might make use of the translation norms of the time. But then again, the very process of breaking down sentences into phrases or syntactic structures and rebuilding them might water down the norms and generate some other SMT intrinsic norms. Studies would be required to determine if this is the case.

If we take a wider view, we can see that MT does indeed use skopos theory. The techniques of domain adaptation and the use of in-domain data imply that the MT developer is thinking about the users of the TT and modifying what the MT system should output accordingly. Translators change their translation strategies in order to adapt the TT to the target culture. Since changing the MT system is costly and can involve much research and effort, the next best approach is used to adapt the TT: use data that closely resembles it. The methods and results are different, but the aim is the same. This implies that linguistics (or its statistical proxy) is not the be-all and end-all in translation, but that external factors have to be taken into account.

One final point to note about the previous discussion is the return to a human. Trying to understand MT from a theoretical perspective forces us to step back from the system itself with its mathematics or rules and into the human realm where we can attribute intelligent phenomena such as intentionality and thought. In the case of MT, it is the system developers and system implementers that have stepped into the shoes of translators and have to carry out the translation. Instead of doing it all themselves, they proxy the work to computers: a set of rules or a parallel corpus takes the place of memory, mathematics and algorithms take the place of translation strategies and procedures. This view rings close to what Pym (2009) describes for the localisation paradigm: splitting what a single translator used to do alone into parts that are performed by different people or even computers. In other words, translation as a distributed system. MT is simply the expression of translation where the majority of those parts are taken over by a computer.

2.3 TRANSLATION PROCESS RESEARCH

Translation Process Research (TPR) is a theoretical branch closely related to descriptive translation studies (Section 2.1.3) that aims to describe how translators translate. It focuses on aspects such as translation units, styles, phases, variation among translators, etc. Essentially, it provides a detailed description of how a translation is carried out by a translator, not just a description of the final text itself.

Christensen (2011) provides an overview of the cognitive grounding for translation and TPR, a brief summary of which now follows. Some aspects of the translation process can be observed directly, but a large part of it occurs in a translator's mind. To understand the translation process better, we need to access the cognitive phenomena that go on in the mind. Risku (2010) notes that cognition is not only a mental affair, the environment plays a direct role in the thought processes. So, understanding translation requires observing what goes on in the mind as well as what goes on in a translator's work environment. Hutchins (2000) proposes the Distributed Cognition Paradigm, in which cognition is distributed across people and artifacts that make humans smarter. In other words, we think with our minds and with our tools, artifacts which can change and shape our mental processes. Thus, when developing tools for translators it becomes relevant to see how they affect their mental processes, and not just the final translation.

In order to gather data on the cognitive processes that go on in the mind, several methodologies exist. Christensen (2011) provides a detailed classification and description of the positive and negative aspects of each. A first major distinction is whether the data is collected after the translation (offline) or as the translation is being carried out (online). Within each group, two further subgroups can be defined. Offline methods can include those that analyse the product (i. e., the translations) or verbal-report data (such as retrospective questionnaires). Online methods can include behaviour observation (eye tracking, keylogging, screen recording, etc.) and verbal-report data (such as think-aloud protocols, where translators say what they are doing while they are doing it). A detailed description of each is outside of the skope of this thesis, but the differences mainly relate to how much each method affects and can potentially alter the process it's trying to observe and how reliable the data gathered is. To overcome the potential downfalls of any one method, triangulation is used. Triangulation (Alves, 2003) means collecting data using more than one of the methods previously described in order to contrast and complement the obtained data, gaining a better view of the overall picture.

Many recent studies have used TPR to study translation and translation tools such as translation memories and Post-Editing. Keylog-

ging and eyetracking (O'Brien, 2009) have been particularly popular. Databases have been created in an effort to standardise the collected data and allow for experiments on large datasets to be carried out, such as the CRITT TPR-DB (Carl, 2012). This thesis will follow in these footsteps at a smaller scale using some of the more basic TPR methods (those that don't require specialised hardware), as will be discussed in Chapter 5.

3 COMPUTER-ASSISTED TRANSLATION

3.1 MAN, MACHINE AND EVERYTHING IN-BETWEEN

Let us place translation on a linear continuum. On one end sits Human Translation (HT), performed by human translators from scratch. However, this does not mean that they don't use computers to search online dictionaries, databases, websites, articles, style guides, help forums and a plethora of other resources. They do, and it is an integral part of the tasks a translator carries out.

On the other end of the continuum sits MT, perfomed by machines without human intervention. However, this is only partly true. The most widespread type of MT systems today, SMT systems, rely on huge corpora of aligned texts previously translated by human translators. Even rule-based systems rely on the human thought put into building the rules. Just like skopos and descriptive theories grounded translation into the real world as an activity carried out by people, away from the abstract equivalence paradigm, the same can be done for MT. As described in Section 2.2, MT does not carry out translation in a platonic world of ideas and *argmax*es, but someone in the real world designs the systems and someone uses them to translate texts for a certain purpose. Thus it is clear that our linear continuum is not as black and white, man vs. machine, as it first seemed.

3.1.1 *Current State of CAT Tools*

Somewhere in the middle of the continuum, closer to HT than MT, sit CAT tools. These provide a wealth of features to help translators carry out their translations. The most widely-used of these tools today is Trados[1]. In 2006 a survey of language professionals reported that 76 % of its respondents used it (Lagoudaki, 2006).

The main feature of CAT tools are Translation Memories (TMs). A TM is a database of a translator's previous translations, segmented into sentences and aligned. When a translator has a new text to translate, each sentence-segment in the ST is string matched against those in the TM database. If a match is found above a pre-determined threshold of similarity, the match is presented to the translator as a suggested translation for the current segment. This tool is especially useful when translating updates to manuals or other similar text types, where the majority of sentences have already been previously translated.

1 http://www.translationzone.com/trados.html

Similarly, glossaries of bilingual terms can be created and function in much the same fashion as a TM, but for smaller units such as words or multi-word units. Recent research has focused on augmenting TMs by including smaller subsentential units, effectively combining them with glossaries. For example, Chuang, Jian, Chang, and Chang (2005) present a method to extract collocations from corpora and build TMs out of them.

Some of the other main features of CAT tools are related to project management, especially for localisation, where texts are frequently non-linear (as discussed in Section 2.1.5) and can come in many formats that are difficult to handle for a translator, such as HTML, XML and others. CAT tools extract the strings to be translated and present them to translators in a consistent interface, usually a two-column window of cells: in one column the ST segmented into sentences and in the other textboxes waiting for the translation of the corresponding segment. Some CAT tools also enable easy collaboration between teams of translators, allowing them to work on a shared TM and glossary hosted on a server.

Recently, research into CAT tools has picked up. For example, the open-source tool MateCat (Federico et al., 2014) has been developed in an effort to integrate MT and CAT tools, moving futher towards setting PE as the norm.

3.1.2 *Problems With CAT Tools*

Despite the many advantages of CAT tools presented in the previous section, numerous studies have identified problems, especially with regards to the way translators translate when using them and how it effects their mental processes. Christensen (2011) provides an overview of a number of studies into TM usage and their effects on translator cognitive processes, some of which are discussed below.

Dragsted (2004, 2006) found that translators don't mentally segment into sentences, but rather tend to use phrases and clauses. This clashes with the sentence segmentation that TM forces. A further finding was that translators using a TM tended to revise sentence by sentence as they go along rather than carrying out a whole global revision at the end as they would when not using one.

Christensen and Schjoldager (2011) investigated the effect of TM on students right after they used CAT tools for the first time. Students reported that the technology was useful but also deceptive, since they felt they lost control of the process, lost track of the aim of the translation, etc. They also reported that translation become mechanical and much less creative and functional.

Moving away from the mental process of translators and looking at the translations themselves, Jiménez-Crespo (2010) carried out a contrastive analysis of web texts localised using a TM and texts sponta-

neously produced in the source language. He found that differences in the text superstructure are present, indicating the source text structure was being closely replicated. Texts translated with a TM also showed lower levels of lexical and typographic consistency. However, he does note that TM is probably only one of many factors that combined to produce the observed differences.

3.2 POST-EDITING MACHINE TRANSLATION

Coming back to the continuum introduced at the beginning of this chapter, Post-Editing (PE) sits close to the MT end. Post-Editing consists in having translators edit a translation produced by an MT system in order to improve its quality. Recently, PE has gained importance both in research and in industry usage. Carl, Gutermuth, and Hansen-Schirra (2015) claim that the recent surge in interest is due to the rise in demand for translation and the inability for human translators to keep up with this demand.

Carl, Gutermuth, and Hansen-Schirra (2015) propose different types and quality levels for post-editing, ranging from light PE aimed at changing only what is necessary for text comprehension, and full PE aimed at bringing MT output to the quality produced by a human translator. However, as the authors discuss, this has an impact on translators as it is difficult to ask them to produce a translation they consider to be of low quality, generating a negative attitude towards PE. Most research into PE revolves around the light variant, with little attention being paid to "full PE".

Carl, Gutermuth, and Hansen-Schirra (2015) also conducted a study with translators, revealing that 83 % of participants would have preferred to translate from scratch rather than use PE. Some authors have argued for the need to teach PE as a separate skill in order to familiarise students with the workflow and expectations (O'Brien, 2002). It is not clear, however, if the approach of adapting translators to the tools is a better way forward than adapting the tools to the translators.

Part II

A UNIFIED METHODOLOGY

4

CAT RESEARCH METHODOLOGY

Developments in the field of CAT such as those mentioned in the previous chapter have been created without an extensive study into the potential effects they have on translators. While reuse, speed and — increasingly — reduced cognitive effort are claimed as advantages of each new improvement, studies have brought to light the not-so-obvious influences these tools have on the translation process, on translators and on translations themselves. Given that TM and CAT tools in general are now an essential part of a translator's environment (O'Brien, 2012), it is probably time to think through what approach to take when developing further improvements.

A common theme of the CAT tools discussed in the previous chapter is that they weigh down human translation with their own inherent limitations and simplifications. As seen in Section 2.2, the fact that MT is basically stuck in the equivalence paradigm can be seen as both a limitation and simplification: trying to include all the external variables into the models used by SMT systems would most likely make the problem intractable. Hardmeier (2014) elaborates on translation theories and some of the simplifications SMT adopted and still adopts to solve the translation problem computationally. Hardmeier (2014) discusses two notable examples: the existence of some kind of equivalence between the ST and the TT, which enables the notion of word alignments between a ST and its TT; and translating only one sentence at a time without considering the wider textual context. These design limitations are now burdened onto translators through PE and TMs (even more so now that research into unifying TMs and MT is picking up).

It is of no surprise that translators tend to dislike using these tools, given that they make their work mechanical and less creative (Christensen & Schjoldager, 2011). Taking a finer-grained look at PE, the very term post-edit seems to convey the idea that the hard work is already done by the system, and just some cosmetic changes are required.

Evaluation of CAT tools is an essential point that needs addressing. Currently, many studies draw conclusions simply based on speed and cognitive effort measured as words processed per unit of time (Carl, Gutermuth, & Hansen-Schirra, 2015). I argue that not all effort is created equal. Translators carry out various kinds of tasks when translating, some which require skills aquired after years of practice and study. It is natural that translating a play on words, a complicated syntactical structure, an unclear passage in the ST, etc. requires more

effort and time than other tasks. These are precisely some of the tasks human translators should be carrying out and are trained to carry out. Offloading key tasks such as word and structure selection to an MT engine and using qualified translators to "clean up" the output is a questionable way forward. As was discussed in Chapter 3, this leads to negative attitudes towards PE.

I propose a new approach to the development of CAT tools that puts translators first from the outset. The methodology consists of the following steps:

1. *Gather automation candidates:* theoretical work carried out in TS and translators themselves through surveys are a valuable source of candidate taskes to be automated. They can provide hints as to which tasks within the translation process are the most tedious and can potentially be carried out without human intervention.

2. *Select candidates addressable by* NLP*:* from the candidates gathered above, experts in computational linguistics can select the ones that can be resonably carried out with NLP tools. Other areas that plain software engineering can deal with could also be identified.

3. *Develop a tool or prototype:* design and implement a system to carry out the selected task. Emphasis is placed on prototype: at this stage, a full implementation may not be required as long as it can be reliably given to translators to use in the context of an experiment. Prototyping allows the tool development process to go much faster, quickly discarding bad ideas and gathering key feedback to improve the prototype once it is time to implement a full system. It could be argued that any difficulties in implementing a full system have not been caught. But in that case the design and efectiveness of the tool has been tested in a real setting and it justifies more resources being spent on overcoming the implementation difficulties. The opposite case, a good system that doesn't mix well with translators, is undesirable.

4. *Intrinsic tool evaluation:* depending on the nature of the tool and whether it is a full tool or a prototype, an intrinsic evaluation should be carried out. At the very least, it should show acceptable performance in the experimental setup described in the next step.

5. *Extrinsic evaluation:* an experimental setup is designed for translators to use the tool. Usage data is obtained by means of triangulation of data-gathering methods developed in TPR. A good mix of methods (online and offline, verbal-report, product analysis and process monitoring) is preferable. The data gathered

should then allow evaluation of the tool's effects on the whole translation process, including translators. The aim will be to answer the question "Does this tool help translators carry out their work?".

The previous methodology should generate tools which do not force limitations onto translators and negatively affect their normal workflow. It is important to stress that the proposed methodology is designed for CAT tool development. The idea of a prototype forces the researcher to not go down the rabbit whole of developing a perfect system. It instead forces them to take a step back and spend less time on the tool itself and more on how the tool is used and if it makes sense for translators. This approach is clearly not ideal for the development of a parsing algorithm, for example. CAT is an eminently practical affair and the development of CAT tools should reflect this fact.

5 EXPERIMENT OVERVIEW

5.1 GENERAL HYPOTHESIS AND EVALUATION

As stated in Chapter 1, this thesis formulates the following hypothesis:

> HYPOTHESIS: The methodology laid out in Chapter 4 will allow CAT tools to be developed and evaluated based on how well they perform intrinsically as well as extrinsically with translators.

In order to evaluate the methodology, a case study was used. Specifically, a CAT tool called StyleCheck was designed and evaluated. Then, a metaevaluation of the methodology is carried out. To sustain the hypothesis, the methodology should:

A. Allow for a CAT tool to be developed and evaluated with translators in mind. Specifically, it should show effects of the tool on the whole translation process (product, process and agent; translation, translating and translator).

B. Provide tools that generate sufficient data for the tool to be evaluated as laid out before.

5.2 ONLINE EXPERIMENT

StyleCheck was developed for the case study, a tool to check if style guide rules are applied in a text and give translators a suggestion if they aren't (see Chapter 6 for details). A special website was created to allow StyleCheck to be evaluated by translators. This allowed participants to work from their own computers in a setting familiar to them. The Spanish Wikipedia style guide was chosen and three text fragments were selected to be translated from English into Spanish. Three setups in which to translate the text were devised:

1. *Translation from scratch* (SCRATCH): The text was translated from scratch without any CAT tool assistance. Translators were free to translate as they normally would, including searching on the Internet and in databases for translations, finding more information about the text topic, etc.

2. *Post-Editing* (PE): A machine translation of the text provided by Google Translate[1] was given to participants. So as to avoid negative connotations that Google Translate may have for translators, they were simply told the suggestions came from an MT engine without specifying which one. Carl, Gutermuth, and Hansen-Schirra (2015) provided participants in their study with a set of guidelines for how to post-edit. No such guidelines were offered in this experiment, instead opting to give translators freedom to use the suggestions as much or as little as they wanted.

3. *Translation using StyleCheck* (STYLE): In this setting, when a participant finishes translating a sentence and moves to the next, the previous sentence is checked using StyleCheck and any returned style hints are presented to the participant. Translators can then choose to follow the style hint or ignore it.

In order to obtain varied data, the text and setup combinations were shuffled. This created six combinations that ensured each text was translated twice in each setup. Thus, the minimum number of participants required for the experiment was six.

Participants were gathered through the Internet. A call for participation in the experiment was posted in various translation-related Facebook groups whose participants are mostly translators or translation students. Participants were asked to contact an email address to request a unique participation link. It was hoped that this approach would limit the number of participants who abandoned the experiment half-way through, as they had implicitely commited to carrying it out. This was especially desirable given that each participant was given a different text-setup order and all possible combinations needed to be covered with the minimum number of participants. If the website had generated the order and Participant ID automatically, a high rate of abandonment or curious people clicking through a couple of pages may have ended with results not covering all possible text combinations and being skewed.

Participants did not receive payment or other kinds of compensation for their participation in the experiment due to the lack of funding available for this thesis. This could influence the results as translators may be inclined to put less effort into the translations if they don't receive any compensation. Ideally, experiments with translators should include payment in line with market prices so that they feel they are translating as they would with a normal client.

Following the proposed methodology, the StyleCheck tool is evaluated using TPR methodologies. Despite many advanced monitoring tools such as eye-trackers and screen recording being used today, this thesis takes a simpler approach. The online website format used for the experiment limits the hardware capabilities to participants' own

1 https://translate.google.com/

computers and limits the software to what can be acheived with a small server and web browser combination.

Both online and offline methodologies were used. The online data recorded is detailed in Section 7.5 and was collected on the server. Offline verbal report data was also collected through the use of retrospective questionnaires. Offline verbalisation does not affect the task and subject's mental processess as much as online methods do (for example, think-aloud protocols). The downside of verbalisation is that if the process to be observed (e.g. PE or using a TM) has become routine, the steps to perform no longer live in the short-term memory and are thus not accessible through verbalisation. This can have an impact if the participants have extensive previous experience with the task to be carried out.

Finally, the collected data was analysed and contrasted in order to see if the hypotheses could be strengthened.

5.3 CASE STUDY: WIKIPEDIA

Wikipedia was chosen as the subject for the experiment case study, since it contains freely available texts and also has style guides for various languages. First, an imaginary translation brief was created so that participants could use it as a guide for their translation choices:

> Wikipedia set up a crowdfunding campaign in an effort to pay for some professional translations of articles very popular in different regions of the world. After a huge success, they invited applications from translators. You applied and have now been selected to translate three articles from English into Spanish. Wikipedia has reminded you of the global nature and scale of the project and has provided a link to its style guide in order to help you translate.

The brief included a link to the Spanish Wikipedia Style Guide. Clicks on this link by participants were logged (see Section 7.5). The three texts used for the experiments can be found in Appendix C. After an in-depth look at the style guide, the texts were chosen following three criteria:

- The text subject matter should not be specialized.
- Various style rules could apply to the text.
- The text should not present major translation problems that would require a lot of time and research to solve.

The texts will be referred to as AI, CHARLOTTE and GARFIELD in the rest of this thesis.

6 STYLECHECK: A TOOL TO APPLY STYLE GUIDES

6.1 STYLE GUIDES

Following the proposed methodology, translation theory was used as a basis to develop a CAT tool that implements a Style Guide (SG). SGs are collections of rules that serve to standardise linguistic content (Vidal Hussey, 2013). Within a language, many different ways of writing or expressing an idea exist. From the wide range of possibilities, dialects choose a subset, as do specific domains, genres, language academies and even individuals. Style guides aim to define a subset of language for use in a specific domain (such as style manuals for scientific writing) or most commonly for a specific entity (Vidal Hussey, 2013). In the latter type we can place SGs from the European Union, from Wikipedia or from newspapers and other publications. The idea is to guarantee that texts written by different authors conform to a set of rules so that they can be recognised as coming from a single entity, creating a specific "voice" of sorts.

Style guides can be more or less restrictive, i. e. aim to cover a larger or smaller subset of linguistic phenomena that need to be standardised. Controlled Natural Languages can be considered an extreme form of SGs where basic syntax and constructions of a language are defined. Style guides tend to focus on formal aspects (such as punctuation and formatting), lexicon, specific linguistic constructions, etc.

Style guides are essentially made up of rules. The structure of these rules is to identify a specific linguistic element (word, punctuation, structure, meaning, etc.) and offer the one or more ways in which it should be expressed, often giving examples of the ways in which it should not be expressed (Vidal Hussey, 2013). Thus, to apply style guide rules digitally requires a system that can detect the specific linguistic element and is able to transform it (Vidal Hussey, 2013).

It is worth mentioning that many style guides also include information not directly related to linguistic standardisation, such as legal considerations for journalists, encyclopaedic knowledge about a certain topic, how to use company software or insert certain symbols into a Word document, etc. This evidences the informal nature of style guides and the multiple aims they sometimes try to serve.

Style guides are an important skill a translator should learn as they are widely used when translating (Washbourne, 2012). A problem for translators is that they may have to handle many different style guides, each for a specific client. In this context, it is easy to mix-up the rules or simply not bother reading through a large style guide for

a small job: the time that would be invested in learning the guide's rules would far exceed the time spent on the translation itself. The proof-reader is then charged with ensuring style guide rules are applied during the revision phase.

StyleCheck aims to solve the problem. As translators work on a text, they will be offered suggestions when rules contained in a style guide haven't been applied. Then, they will be able to immediately solve the problem by applying them, which will reduce the time needed for proof-reading and the time spent learning a style guide.

6.2 GRAMMATICAL FRAMEWORK

StyleCheck is implemented using Grammatical Framework (GF) (Ranta, 2011). GF is a programming language and grammar formalism used to build multilingual grammar applications. GF draws heavily from functional programming and from type theory. The real strength of GF is the Resource Grammar Library (RGL) (Ranta, 2009), which implements basic morphology and syntax for a number of languages. The RGL provides a common, high-level API to use the syntax, allowing for fast multilingual application development. Relevant for StyleCheck is GF's tried and tested use in the field of CNL (Angelov & Ranta, 2009; Kaljurand & Kuhn, 2013; Ranta, 2014).

Grammars in GF are split into two parts: the abstract syntax and the concrete syntax. The abstract syntax is intended to represent semantics, while the concrete syntax (one for each language) links the abstract syntax to a particular string representation. More specifically, functions defined in the abstract syntax are mapped to linearizations in the concrete syntax.

As an example, the abstract syntax funcion

```
fun AvoidSeasons_Hint : StyleHint ;
```

builds a record of type `StyleHint`. In the Spanish concrete syntax, this would be realised as:

```
lin AvoidSeasons_Hint = mkStyleHint "verano" "Don't use seasons";
```

where a `StyleHint` is a record built out of a season "verano" and the suggestion "Don't use seasons". When "verano" is encountered during parsing, a `StyleHint` is built and mapped to the abstract syntax function `AvoidSeasons_Hint`. Then, from the abstract syntax tree representation a different concrete syntax can take the abstract function that was matched in the text and select from it only the suggestion text, which can then be output to the user.

This is a simplified view of how StyleCheck is designed, but it serves to illustrate the main idea: a text is parsed and any detected style rules are mapped to a style suggestion.

6.3 STYLECHECK

6.3.1 *Development Procedure*

The StyleCheck grammar was developed as follows. First, from each text to be used in the online experiment all Wikipedia style guide rules that could be applied to it were identified. Since the style guide refers to the Diccionario Panhispánico de Dudas (Española & Academias de la Lengua Española, 2005) for aspects not specifically mentioned in the style guide, it too was checked. The obtained rules were then classified as rules, hints or lookups in the following manner:

- *Style rules:* linguistic elements that have more than one possible forms, out of which one or more forms are approved by the style guide. For example, the correct spelling in Spanish of the capital of UAE is "Abu Dabi", not "Abu Dhabi" as in English.
- *Style hints:* linguistic elements whose approved form is not self-evident and requires reasoning. Through the use of a suggestion, the translator can be made aware of the issue and decide whether action needs to be taken or not. For example, the Wikipedia style guide recommends not temporally locating events using only seasons, as the months they refer to differ in the Northern and Southern hemisphere of the Earth. In this case, all mentions of a season would trigger a suggestion for the translator to check this aspect.
- *Style lookups:* the third possibility that was contemplated were style lookups. For some elements that have many possibile uses, such as full stops, commas or slashes, showing a suggestion every time they appear as would be the case with a style hint was deemed to be too annoying. Instead, style lookups would allow the translator to select one of these common elements and be taken shown the full range of possibilities. Thus, the translator would only lookup these elements on-demand, when he or she would consider it necessary.

Given that only one rule was classified as a style lookup (the use of the forward slash symbol), it was decided to apply it as a style hint instead and leave the lookup framework at the theoretical design stage to be implemented in future work. The forward slash only appeared once in one of the texts, so it didn't risk being annoying as would be the case if it were a comma or a full stop.

Rules were implemented as a record with fields for optional forms, approved forms and the hint (Listing 1). Style hints and lookups contained all the fields in a rule, except for the `approved` field. All of these fields were of type `Chunk`, which is simply defined as a string.

Listing 1: Record structure for the rule types

```
1 StyleRule = {approved,options,hint : Chunk} ;
2 StyleHint = {options,hint : Chunk} ;
3 StyleLookup = {options,hint : Chunk} ;
```

The main problem with developing style rules and hints is identifying the range of options that can be encountered in a text. Not only is it necessary to think about what possible forms a rule can manifest itself into, but misspellings of these forms also need to be taken into account. In order to help in this effort, several helper functions were defined which generated options on the basis of including full stops or not (to create the possible misspellings for ordinal abbreviations in Spanish: 7.º and 7º), including a space between compounds or not, etc. The `approved` and `options` themselves were implemented either as strings or using types and functions defined in the Spanish RGL, allowing for the full morphological forms to be matched.

Lastly, we turn to the hint format. Vidal Hussey (2013) found that the rules contained in SGs come in all shapes and sizes. Some are included as part of a text and are written in a narrative style, others are included as entries in a dictionary format referring to a specific keyword. The Wikipedia style guide takes the narrative approach, justifying why many of the decisions are made and explaining a lot of the concepts. So as to not introduce another variable to analyse, for this thesis the hints were taken as-is or with very minimal editing from the guide and shown directly to the translator. This generated suggestions that sometimes spanned several lines of text. A better approach would have been to rework the suggestions into one-liners apt for a quick glance, and provide an option to further expand on the descriptions or reasoning behind the rules if desired by the translator.

6.3.2 *Grammar Structure*

Despite the small scale of the grammar and the few rules that were implemented, the full file and module structure of a larger grammar was used in order to show how a full system would be implemented. Figure 1 represents how the full grammar was built.

Code for the three main modules `Wiki`, `WikiMaster` and `WikiHint` can be found in Appendix D

First, we will describe the main grammar components. This subset of the full grammar is shown in Figure 2. The main abstract syntax module is `Wiki`, which contains a function for each of the rules (Listing 2). This abstract syntax has three corresponding concrete syntax modules: `WikiMaster`, `WikiHint` and `WikiEdit`. `WikiMaster` (Listing 3) contains the main parsing mechanics. In it, each abstract syntax function is linearised into the required type: `StyleRule`, `StyleHint` or `StyleLookup`:

- These three types are defined in the abstract syntax `StyleCatAbs` and in the concrete `StyleCat`.

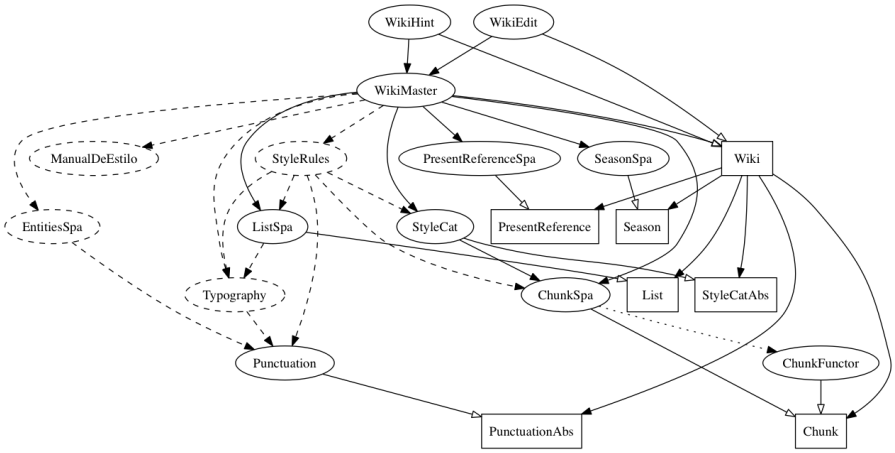


Figure 1: Full module structure and dependencies that make up StyleCheck

- Convenience functions to build these types are provided in the resource module `StyleRules`.
- To build the options, `WikiMaster` uses the necessary components from the Spanish RGL (not shown in Figure 2).
- The text for the style hints is contained in `ManualDeEstilo`.

Once the required style types are built, `WikiMaster` builds `Chunk` types out of them. The idea is that when parsing a sentence, once or more `Chunks` (each corresponding to a style rule, hint or lookup) are identified, appended together, and output as the result of the parse. To acheive this, StyleCheck makes use of the chunking types and functions used in the Wide-Coverage Translator being built upon GF (Ranta, 2014). The abstract syntax for these types is `Chunk`, and the concrete is `ChunkSpa`. Finally, the top level type, a `Phr`, is built out of one or more `Chunks`.

To sum up, the parsing mechanics are as follows:

1. A style rule is matched and a style rule, hint or lookup type is built.
2. `WikiMaster` selects the `options` from the style record and builds a `Chunk` out of them.
3. Through the functions included in `ChunkSpa`, a `Phr` type is built, which is the top level category in the grammar.
4. The tree constructed until now is linearised using `WikiHint`. This concrete syntax selects the `hint` field from the style type record, resulting in the hint text being output.
5. A simple regex extracts all the hints from the output (hints are delimited with `|||` to make this task easy). Any non-matches will simply return the original token.

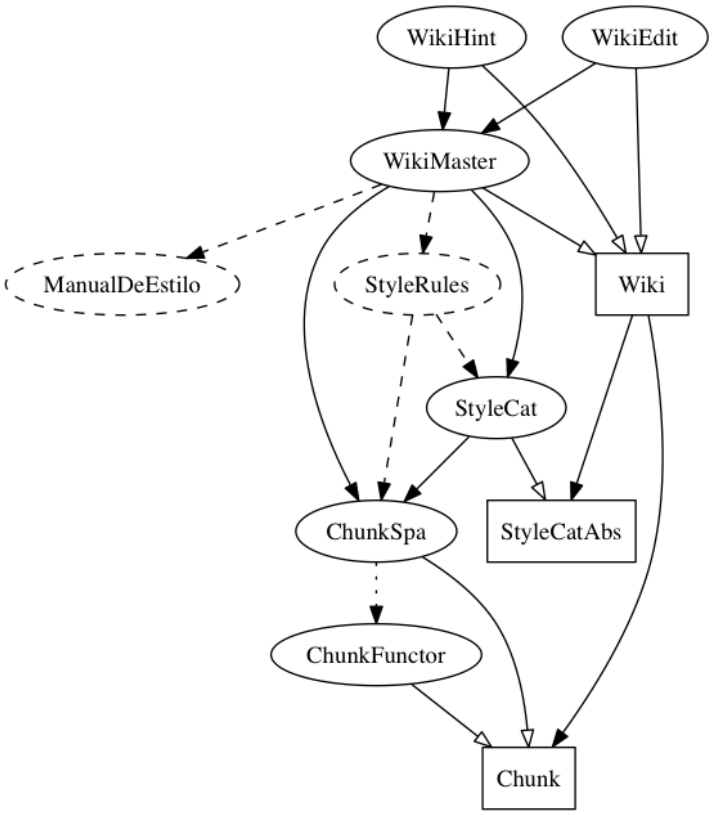


Figure 2: Main StyleCheck modules and dependencies

The additional modules shown in Figure 1, such as `ListSpa` or `Typography` ease the task of writing style rules. They mainly contain functions and opers related to a specific rule in StyleCheck.

Finally, there is one extra concrete syntax that isn't used in this thesis. The `WikiEdit` shows how an automatic post-editing tool would work. It simply selects the `approved` field from the style record instead of the hint for `StyleRules` (the only types that have an approved field). This kind of funtionality for CAT tools needs further thought and development, since it could be very confusing for translators to find parts of their texts changed without warning.

6.4 INTRINSIC EVALUATION

In order to develop the grammar, the possible options for each rule had to be thought out. A test sentence was created for each option, together forming a small corpus. The grammar was developed so that it would generate the correct output on the small corpus of options. This was evaluated by having GF parse the sentences and check that the grammar brings up the correct style hint.

Intrinsic evaluation as described will perform well on the corpus, but some cases can fall through the net. Looking into the final trans-

lations carried out by the experiment's participants, two cases were found of options that the grammar did not detect:

- Rule 2.B (Listing 2, Listing 3) states that temporal references should not use the moment of enunciation as reference point. For example, in "Her most recent work", *recent* refers to the moment of enunciation. Future readers may consider *recent* as a different time period than the writer, so it should be avoided. Even though StyleCheck was developed with several adverbial time markers that refer to the moment of enunciation (*recientemente*, *hoy*, *ahora*, etc.), one participant used an adjective form instead ("la más reciente"), which was not in the grammar and did not generate a hint.
- Another simpler case is that of the USA. The acceptable forms in Spanish are EE. UU. (with a space) or EUA. Several non-approved options were generated, such as EE.UU. (without a space) or USA. One participant, however, used the form EEUU, which was not in the grammar and would not have triggered a style hint.

The advantage of having built a prototype is once again clear. The previous issues can be solved straight away and directly integrated into a future version of StyleCheck.

7 EXPERIMENT WEBSITE

7.1 EXPERIMENT WALKTHROUGH

A detailed walkthough of the experiments with screenshots is presented in Appendix A. The following subsections provide a more detailed description of the technical side of the server setup (Section 7.2), how participants found out about the experiment (Section 7.3), the StyleCheck user interface (Section 7.4) and the data that was logged (Section 7.5).

7.2 SERVER SETUP

The website where participants carried out the translation tasks was created specifically for this experiment. The website was built using the Flask[1] (v. 10.1) web framework and Python (3.4.3) from a virtual environment to help manage package dependencies. The `html` and `css` code was put together using Bootstrap[2]. The dynamic parts of the user interface were written in `javascript` using the jQuery[3] library. Using these frameworks, the website was responsive and thus able to be used on mobile devices, although not recommended for this experiment. Despite this recommendation, one participant reported carrying out the experiment on an iPad.

The server ran Ubuntu 12.04 and was set up with nginx[4] (v. 1.9.0) as the front-end reverse proxy and uWSGI[5] (v. 2.0.10) as the application server communicating with Flask through the WSGI protocol[6]. The server had 16 GB of RAM, 8 CPUs and 40 GB of SSD disk.

Note: Due to the embedding of the Google survey form, a large blank space appears in Figure 17. Please see Section 7.2 for further explanation.

The questionnaires were made using Google Forms[7] and the results downloaded and analysed as `.csv` files. Forms created with Google Forms can either be answered by following a link or they can be embedded into a website. For this experiment, the latter option was chosen so as to be less jarring for participants. However, forms can only be embedded into an **`<iframe>`** element. Since some questionnaire pages contained more questions than others, some pages included a lot of blank space in the **`<iframe>`** (see Figure 17) while others with more questions had to be scrolled.

1 http://flask.pocoo.org/
2 http://getbootstrap.com/
3 http://jquery.com/
4 http://nginx.org/
5 https://github.com/unbit/uwsgi
6 https://www.python.org/dev/peps/pep-3333/
7 https://www.google.com/forms/about/

7.3 EXPERIMENT ACCESS

To access the experiment, participants were requested to contact an email address provided in the Call for Participation announcement. Then, they were provided with a unique link from which to access the experiment. The link URL included a unique Participant ID, which identified the participant throughout the experiment, and a unique text-setup combination. This enabled the experiment to be carried out with all the possible number of combinations for texts and setups. For example, the URL `www.dvh.io/main/420/123` indicates that the Participant ID is 420, Text 1 should be used in the SCRATCH setup, Text 2 in the PE setup and Text 3 in the STYLE setup.

7.4 STYLECHECK USER INTERFACE

The third setup interacted dynamically with the server to show participants style hints as they translated. Every time a `<textarea>` where translations where input lost focus, the contents were sent to the server, analysed using StyleCheck, and any style rules parsed sent to the participant's browser. Before sending them, the hints were checked for duplicates (some hints could be generated multiple times in a parse) and if any were found only one instance of the rule was returned. The rules then showed up in a red panel below the corresponding sentence that was parsed (Figure 3).

Figure 3: Hints being shown to the participants after analysing their translations with StyleCheck.

In some cases, due to the length and number of the hints, the panels could take up a lot of space. Depending on the size of the participant's display and the size of the browser window, the hints could even be rendered outside the visible part of the screen, requiring participants to scroll (Figure 4).

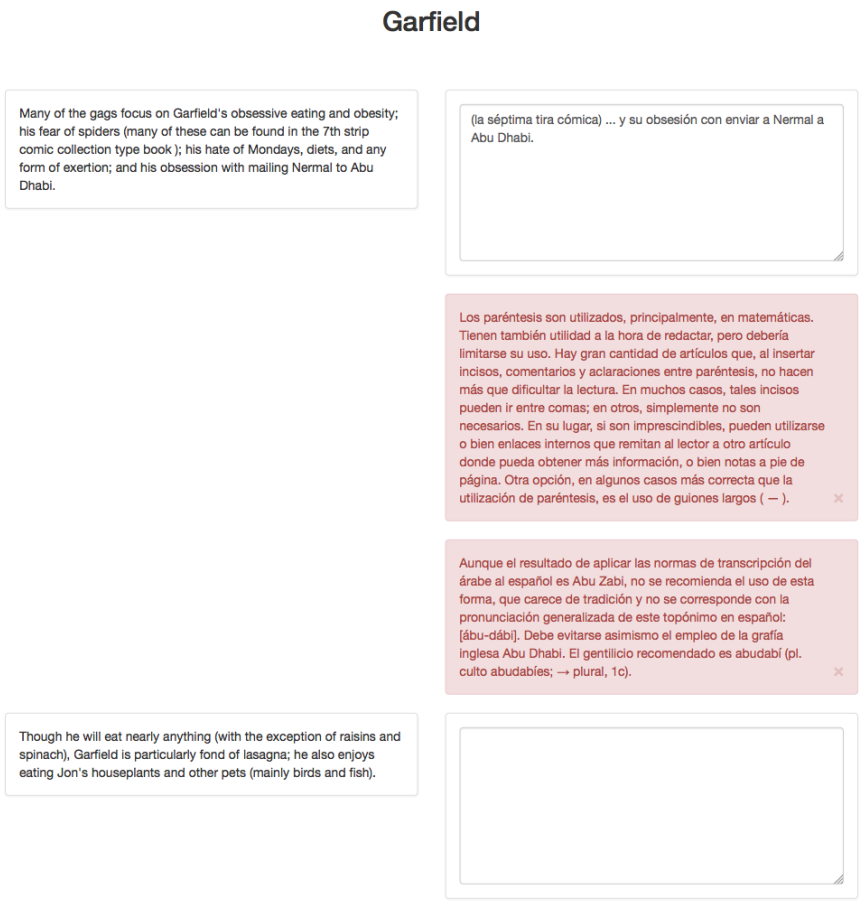


Figure 4: Two long hints being shown taking up a large part of the screen.

7.5 LOGGING

Different kinds of logging, in addition to saving the final translations, were set up in the server:

- *Style guide click*: When a participant clicked on the link to Wikipedia's style guide, the page the participant was on (the translation brief is shown on four different pages, see Appendix A) and the click time was registered.
- *Translation information*: The website saved the actual translation, as well as the start and end time for each setup.
- *Hints*: For the STYLE setup, the contents of the `<textarea>` element, the time and any style hints returned were logged.

8 QUESTIONNAIRES

For this study, two questionnaires were created. The initial questionnaire was completed by participants before carrying out the translation tasks, while the final questionnaire was filled in after.

Note: For screenshots of the actual questionnaire, please see Appendix B.

The questionnaires contained both open and closed questions. An overview of the questions and the research objectives each one aimed to cover are provided in the following subsections. Each question is given a number, although the actual questionnaire did not number the questions so that it didn't seem too long to participants. For this same reason, each questionnaire was presented split into pages and a progress bar shown at the bottom of each page. Options for the closed questions were randomly shuffled.

Given that the majority of participants were likely to be native Spanish speakers and non-native English speakers, the questions were written in English but participants could answer the longer questions in either Spanish or English. Since they would be translating from English, a certain level of knowledge of English was presupposed and it wasn't deemed necessary to offer the questions also in Spanish.

8.1 INITIAL QUESTIONNAIRE

This questionnaire gathered basic demographic data on the participants, as well as their previous experience with translation and various translation-related tasks. It was also exploratory in nature in that it tried to discover what tasks translators would like to be automated, as well as their opinions on MT and it's relation to CAT.

The specific research aims for the initial questionnaire and the questions related to them are the following:

Note: numbers in brackets refer to the specific questions in the initial questionnaire. Please see Appendix B for the full questions.

- Basic participant demographics (age, gender, occupation, languages) [Questions 2, 3, 4, 5, 6]
- Participant's experience and training as translators [Questions 7, 8, 9, 10]
- Participant's experience with specific translation-related tasks [Question 11]
- Participant's experience with specific translation-related tools [Questions 18, 19, 20]
- What do translators themselves consider translation to be? [Questions 1, 15, 16, 17]

- What are the most important considerations when translating? [Questions 15, 16, 17]
- What tasks carried out during translation are candidates for automation? [Questions 12, 13, 14]
- What attitudes do translators have towards machine translation? [Questions 21, 22, 23]

8.2 FINAL QUESTIONNAIRE

The final questionnaire asked translators to compare the three translation tasks performed and asked specific details about each individual task. Following Krings (2005), this questionnaire can be categorised as a retrospective questionnaire (related to a specific task) and as online (the researcher was not present while the questions were answered).

Christensen (2011) notes that a downside to retrospective questionnaires is the risk that subjects might make up explanations since they have to retrieve information from long-term memory. This study tried to partially offset this concern by providing participants with the full ST (and MT suggestion for the second task) directly above the final questionnaire, so that they could more easily remember the texts and answer the questions (see Figure 22). The participant's own translations were not provided so that they would not get frustrated at seeing any possible mistakes they had made.

Note: numbers in brackets refer to the specific questions in the final questionnaire. Please see Appendix B for the full questions.

The specific research aims for the final questionnaire and the questions related to them are described in the following list:

- Which setup was preferred by the translators? [Questions 14, 15, 22, 23, 32, 33]
- Which factors made a text easy or difficult to translate? [Question 3, 4, 5, 6]
- Which setup helped translators produce a better translation (perceived or real)? [Question 15, 23, 33]
- Did they perceive any differences regarding how in control they were of their own final translations between the setups? [Questions 7, 8, 9, 10]
- Did their perceived translation speed vary significantly from their real translation speed? [Questions 11, 17, 25]
- Were the style hints useful for participants? [Question 28, 29, 30, 31, 32, 33]
- Did participants find the MT suggestions useful? [Question 20, 21, 22]

Some of the questions were exploratory in nature and could be categorised into various of the previous aims depending on the participant's answer. Very general and open questions were asked to encourage participants to come up with their own ideas about the three setups. Mainly, these questions were 3, 4, 5, 6, 12, 13, 16, 18, 19, 24, 26, 27, 34 and 35.

Questions 14, 20, 22 and 32 followed those used by Carl, Gutermuth, and Hansen-Schirra (2015). The wording was slightly adapted to fit in with the current study. The question "Overall, how satisfied were you with the task of translating Text X?" (questions 14, 22 and 23) was complemented with "Overall, how satisfied were you with the quality of your final translation of Text X" (questions 15, 23 and 33) to separate the performance of the actual task from its output. The single question offered by Carl, Gutermuth, and Hansen-Schirra (2015) could be ambiguous and make participants hesitate between liking the task but not being satisfied with its final result.

A possibility that was considered was to include the System Usability Scale (SUS) (Brooke, 1996) questions into the final questionnaire. The SUS is a widely-used set of 10 questions that measure the usability of a system. Despite the advantages of using a well-known questionnaire, it was decided to not include them. The 10 questions would have had to be provided for each of the three setups to allow for comparison. An extra 30 questions added to the final questionnaire was deemed to make it too long and fatiguing for participants.

Part III

RESULTS

9 RESULTS AND DISCUSSION

9.1 PARTICIPANT DEMOGRAPHICS

Questions are prefixed with a letter indicating which questionnaire they appear in (I for initial and F for final) followed by the question number. For a detailed look at the questions, please refer to Appendix B.

A total of 11 participants answered the Call for Participation (see Section 5.2) and requested a participation link. Of those, one did not attempt to complete the experiment. One further participant completed all experiment parts except filling in the initial questionnaire, so the provided results were not used. Thus, in total 9 participants completed the experiment, which represents 82 % of people who requested to participate. Given that one participant mostly completed the experiment and a technical issue may have caused the initial questionnaire not to be saved, that brings the participation rate up to 91 %. These figures suggest the approach of asking interested people to request a link and implicitly commit to carrying out the experiment (see Section 5.2) was effective.

Genderwise, 4 male (44 %) and 5 female (56 %) participants took part, all in the 20-25 age group. All participants considered themselves native Catalan speakers, while all but one (89 %) considered themselves native Spanish speakers (the odd one out did select Spanish as one of the languages he or she speaks). Other languages spoken by the participants include English (100 %), French (78 %), Spanish (33 %), Japanese (22 %), German (11 %), Italian (11 %), Korean (11 %), Dutch (11 %), Polish (11 %) and Catalan (11 %). As well as the aforementioned participant, the Spanish and Catalan results are due to a participant selecting them in both the native languages and other languages questions.

All participants have studied translation, during either 4 years (67 %) or 5 years (33 %) [Question I.9]. Regarding professional experience as translators, 56 % reported having worked as translators, while 44 % hadn't. Of those that had, 40 % had worked for 2 years as a translator, 40 % for 1 year and 20 % for less than 1 year [Question I.10].

When asked specifically about their professional experience [Question I.11], 56 % reported having translated from scratch; 44 % reported experience with proof-reading, PE and localising; 22 % said terminology management, using CAT tools and preparing documents for internationalisation; and 11 % mentioned transcreation, setting up style guides, MT quality evaluation and subtitling. It is interesting to note that the answers to Question 11 and the previous Questions 9 and 10 don't match up as expected, i. e. some participants who reported not having worked as translators then went on to say they had professional experience translating from scratch, for example. It is possible

that in interpreting the question, participants included internships and odd jobs which they did not fully consider to be "work".

Given the previous descriptions, the participants can be classified as semi-professionals: they are no longer students, but still not as experienced as a long-standing professional.

9.2 CONCEPTIONS OF TRANSLATION

This section describes what the participants think about translation [Questions I.1 and I.12–I.23]. The aim of these questions is two-fold. On the one hand, they aim to provide more context for the tool evaluations the participants provide. On the other hand, they aim to verify if translators are a useful source of suggestions of how to improve CAT tools.

9.2.1 *Translation Tasks*

Participants were asked which tasks they considered boring and which they considered interesting. It is assumed that the boring tasks would be those that a translator would prefer not to do and thus could be considered as a candidate for automation.

Participants considered that researching and learning about a new topic is one of the most interesting tasks in translation, as well as the process of translation itself [Question I.12]. The problem solving aspect was also frequently mentioned, such as solving cultural clashes and difficulties present in the text. Creativity also came up, a topic which can surprise those who consider translation to be an essentially mechanical task but which is frequently discussed in translation theory. As for what participants found boring when translating [Question I.13], 33 % also mentioned topic research. Other aspects considered boring were revising and proofreading, page layout management, terminology research and database management (I interpret this as glossary and translation memory management). One participant even explicitly mentioned PE as a boring task.

The questionnaire asked participants that if they could have a tool to automate any task in translation, which task would it be [Question I.14]. A few mentioned the same tasks they considered boring, such as topic research and proofreading. Interestingly, many of the tools they wished for were those that provide suggestions of various kinds: collocations and expressions, synonyms, translation options in context and problematic structures in the text. This suggests participants wished to remain in control of their translations and have the power to pick an option, rather than being forced an option in the case of TM and PE which they then have to change only if necessary (as most PE guidelines state, see for example Carl, Gutermuth, and Hansen-Schirra (2015)).

Thus, it appears that simple questionnaires asking translators what tasks they would like to be automated is a useful way to come up with CAT tool improvements. It also shows that not all improvements can please everyone: one participant stated they had all the tools they needed to translate while another explicitely mentioned PE as an activity they dislike.

9.2.2 *What Makes a Good Translation*

The question of translation evaluation —what a good and a bad translation is, if they can even be classified as such— is inextricably linked to how translation is conceptualised. Participants were asked what translation meant to them [Question I.1]. To analyse the responses, the translation theory paradigms discussed in Section 2.1 provide a good starting point. Given the freeform answers and open-ended question, it is difficult to classify a participant's views in one paradigm or the other. However, a large number of responses describe translation in a similar fashion to the equivalence paradigm, some nuancing this description with tints of functional priorities and the need to adapt the text to the target culture and function.

To garner a more detailed and nuanced view, participants were asked to rate specific considerations on a scale of 1 to 5, where 1 is very important and 5 is not at all important [Question I.15]. As can be seen in Table 1, all considerations were important to the participants. Grammaticality of the final product was the most important priority, followed by conforming to client requirements. Interestingly, accurately portraying the ST was second to last in the ranking, only more important than applying a style guide. These results seem to indicate that despite their initial answers, participants view translation in line with the skopos/functional paradigm.

When asked which of the considerations they prioritised the most [Question I.16], 56 % indicated providing a fluent/idiomatic translation, followed by providing a grammatically correct translation and accurately portraying the ST with 22 % each. Once again, even though the ST is important, TT considerations are generally more important. Regarding other considerations [Question I.17], one participant mentioned varying his or her priorities depending on the TT function.

9.2.3 *Tools and Resources*

Regarding what tools the participants normally use, 78 % reported using both glossaries and style guides. CAT tool usage was lower, only 57 % stated they regularly use them when translating.

Another interesting aspect were participants' views on MT. The majority of participants were uncertain about its usefulness to translators (67 %), with the rest saying it was useful. Similar views were

CONSIDERATION	MEAN RATING	σ
Providing a grammatically correct translation	1.2	0.44
Conforming to what the client requests in the translation brief	1.4	0.53
Providing a fluent translation/idiomatic translation	1.6	0.73
Using terminology consistently	2.0	1.12
Accurately portraying the source text	2.1	0.78
Applying a style guide consistently	2.2	1.20

Table 1: Mean ratings and standard deviation for the importance of various considerations when translating, with 1 being very important and 5 being not at all important.

gathered when asked if MT is only useful for the general public: 44 % remained neutral, while a third of them disagreed. This could indicate that some of those hesitant to say MT is useful for translators conceded that it does have professional use beyond gisting systems.

The majority of participants did not think MT will replace translators (78 %), the rest being uncertain about it. This could be related to their perceptions of MT quality, as 67 % considered that it does not provide good quality translations today, with 33 % being unsure. The uncertain responses could indicate that the translators are out of touch with current MT research and how MT systems work: 44 % stated it's difficult to understand how MT works and the rest were devided between being unsure (22 %) and saying it is easy to understand (33 %). A large majority (78 %) did state their willingness to learn more about MT.

Finally, all participants considered their translations are better than those provided by MT. When asked why, they responded in a more nuanced way. As one participant put it, "anything beyond grammar and vocabulary is lost". Many mention its lack of adaptability to client and target culture requirements, it can't deal with humour and cultural clashes, etc. In essence, it lacks the "human factor" as some participants put it. This makes the translations "very standardised" and doesn't take style into account. Despite all the downsides, one participant did concede that the results are good enough and fast for certain language pairs.

9.3 TIME AND SPEED

Time is one of the preferred metrics for evaluating PE. A large number of studies present it in the form of speed defined as *words/hour* (Fed-

erico, Cattelan, & Trombetti, 2012), although others use *seconds/word* (Koponen, Aziz, Ramos, & Specia, 2012). For this thesis, *words/hour* was chosen. Two forms of time data were collected: real time spent on each translation by way of logging on the server app and the time translators perceived they spent on each translation by way of the final questionnaire [Questions F.11, F.17 and F.25]. The results were converted to *words/hour* and analysed.

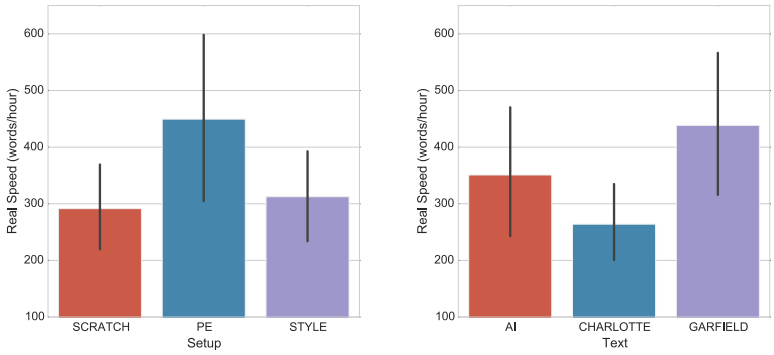
It is also worth noting that this study considers the total translation time for each text from opening the document to saving the final translation. This time could include Internet searches, glossary and other linguistic resource lookup, a quick glance at a phone, a sip of coffee, etc. This is in contrast to the approach taken by Federico, Cattelan, and Trombetti (2012): they only consider segments where the translator is actually "translating". They determine this by setting a processing threshold for the segments included in their analysis (between 30 and 0.5 *seconds/word*) and discard the rest. This threshold ignores many of the tasks a translator carries out and which should be considered part of the translation process, such as terminology lookup and topic documentation. It can often be the case that searching for a single term can take longer than "translating" the rest of the text. Even though the authors specify that their choice of speed as an evaluation metric is intended to correlate with translation cost, the thresholds they set ignore the tasks carried out during translation which can be the most costly in terms of time.

9.3.1 *Setups and Texts*

Both text and setup have an effect on the translation speed. Figure 5a shows that PE is the fastest setup to translate in, with SCRATCH and STYLE performing similarly. Figure 5b shows that text also plays an important role: GARFIELD achieves the highest average speed across all setups, followed by AI and then CHARLOTTE. This lines up with how much participants liked to translate a certain text and how hard they thought it was (this data is detailed in Section 9.4). Digging further into the data, Figure 6 shows that the fastest combination by far is GARFIELD in the PE setup. SCRATCH shows speed variance among texts, while STYLE evens out the speed differences.

9.3.2 *Participants*

Individual variation is a strong factor in speed considerations. As can be seen in Figure 7 and Figure 8, speeds can range between just over 100 to over 700 *words/hour* depending on the translator and the setup. Some participants such as 421, 423, 425 and 427 translate at a fairly constant speed across texts and setups. Others, such as 424 and 429, experience major speed differences across setups. In particular,



(a) Setup (b) Text

Figure 5: Mean speeds per setup and per text. 95 % confidence interval calculated using bootstrap resampling.

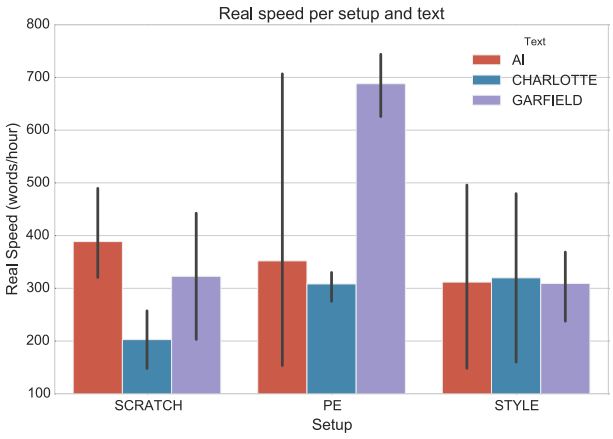


Figure 6: Mean speeds broken down per setup and per text. 95 % confidence interval calculated using bootstrap resampling.

Figure 8 shows these latter two participants experience a major speed-up in the PE setup, which could indicate they only minimally edited the MT output. To confirm this, BLEU (Papineni, Roukos, Ward, & Zhu, 2001) scores were calculated between the MT suggestions and the final post-edited translations. BLEU is usually used to score how well a machine-translated text matches a reference translation, here the metric is used to see how much editing was performed on the MT suggestions by the participants. Figure 9a shows a correlation (Pearson's $r = 0.75$) between the BLEU scores and the time spent post-editing. This correlation is also present in the time participants thought they had spent on the task (Figure 9b, Pearson's $r = 0.8$).

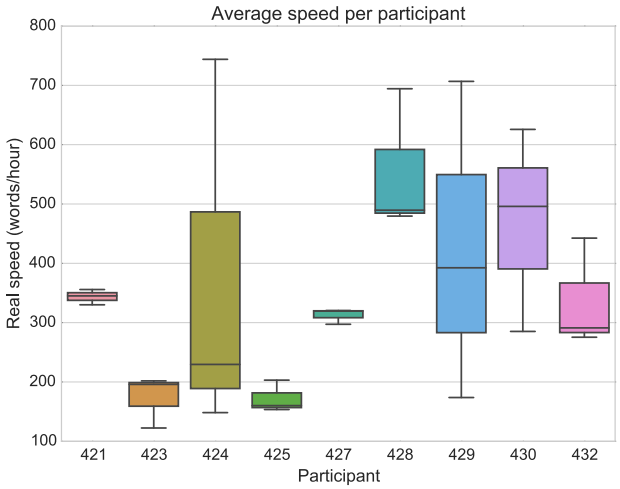


Figure 7: Mean real speeds per participant

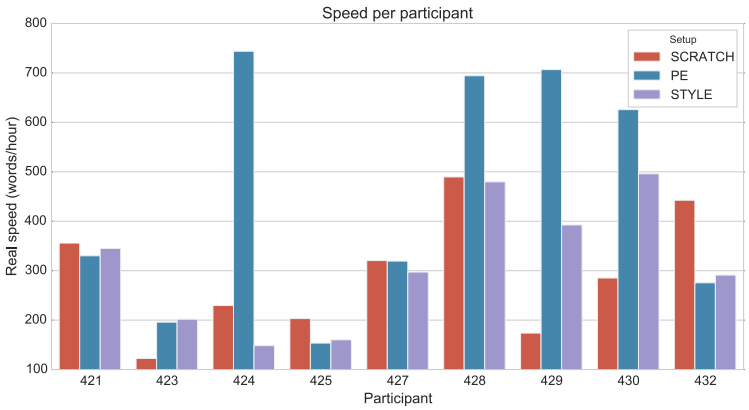
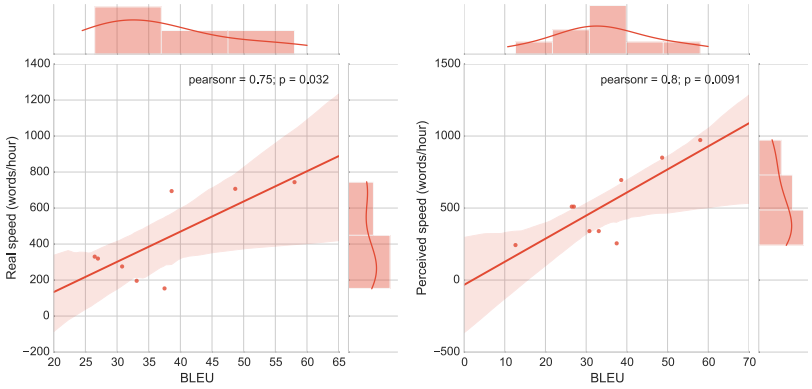


Figure 8: Real speed per participant and setup



(a) Real speed (b) Perceived speed

Figure 9: Correlation between real and perceived speed spent on post-editing and the final translation's BLEU score. Data from one participant was omitted from the real time as it was a clear outlier (including it, the correlation was $r = 0.39$)

Thus, it seems that the PE only acheives an increase in speed when translators minimally edit the MT suggestion. Participants were not given guidelines on how to use the MT suggestions, so they were free to take the approach they preferred. If they feel they need to change the suggestion more, speed differences are negligible or can even result in a slow-down (Figure 10).

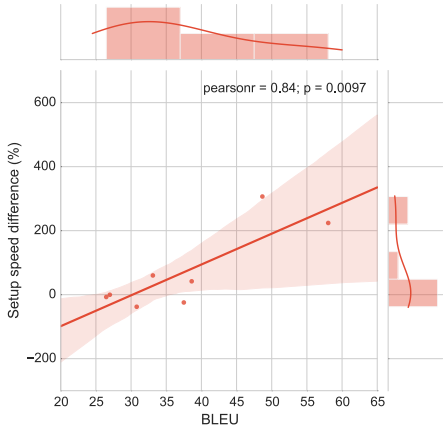


Figure 10: Correlation between the % speed difference between SCRATCH and PE (positive indicates PE was faster) and the final translation's BLEU scores. Data from one participant was omitted from the real speed as it was a clear outlier (including it, the correlation was $r = 0.56$)

9.3.3 *Perceptions of Time*

Perceived speeds have not been studied in previous literature on translation process research, thus there is no background in how to

interpret them. I suggest they could also be considered a measure of effort. A tedious task raises awareness of the time being spent on it and makes people think they're spending more time on it than they actually have, while an easy and enjoyable task can seem faster. The results in this study seem to support this view.

As can be seen in Figure 11, overall the translators thought they translated much faster than they actually did. Digging further into the data, Figure 12 shows that participants felt that their speed between setups was roughly similar, with PE as fastest and STYLE as slowest. This contrasts with the real speeds, which were roughly the same for STYLE and SCRATCH and faster for PE.

In other words, participants were fairly accurate at predicting their time spent in PE, but they felt SCRATCH and STYLE were faster than they actually were. This indicates that they were more aware of the time during PE. Supposing a link between time awareness and effort, this would indicate PE required more effort and was liked less.

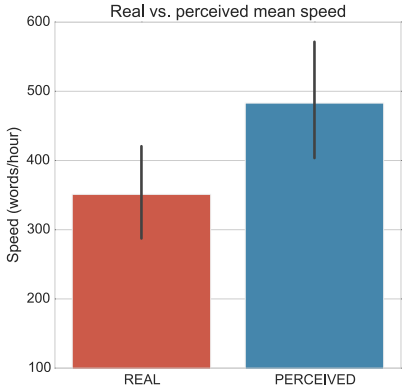


Figure 11: Overall real vs. perceived mean speed. 95 % confidence interval calculated using bootstrap resampling.

As will be discussed in Section 9.4, participants associated like and dislike more strongly with a particular text than with a particular setup. Looking at the speed per text in Figure 12, we see participants perceived CHARLOTTE as being the slowest text to translate, which lines up with it being considered the hardest (Section 9.4.2). AI was both considered the hardest (4 participants) and the second easiest (3 participants), which could explain why its average perceived speed is roughly as fast as GARFIELD, which is considered the easiest translation (5 participants).

9.4 SETUP PREFERENCES

In order to determine which setup the participants preferred, two key questions were asked: whether they would have preferred to

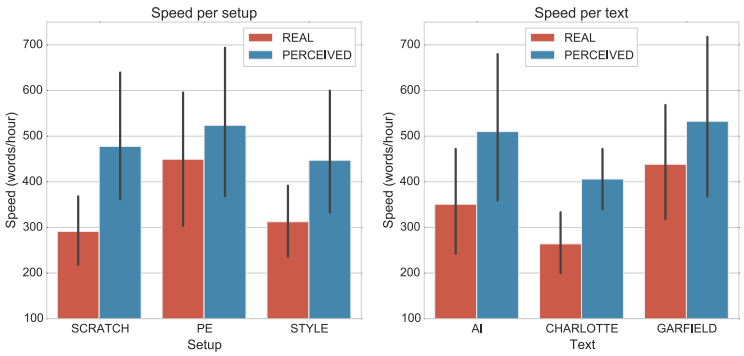


Figure 12: Real vs. perceived speed per setup and per text. 95 % confidence interval calculated using bootstrap resampling.

translate from scratch without the MT suggestion and without the StyleCheck hints [Questions 21 and 31]. 44 % said they would prefer not to have the MT suggestions versus 33 % who would prefer not to have the StyleCheck suggestions. Thus, overall acceptance is higher for StyleCheck than for PE. The acceptability rate for PE was higher among this study's participants than in other studies such as in Carl, Gutermuth, and Hansen-Schirra (2015), where 83 % of participants responded preferring translating from scratch. Carl, Gutermuth, and Hansen-Schirra (2015) does report a higher satisfaction rate for the PE task among students versus professionals. Since this study's participants had limited experience as translators ($<$ 2 years), they could be considered closer to the student category than the professional category. This could explain the higher acceptability rate.

9.4.1 *Satisfaction*

Participants were asked to rate their satisfaction both for the task and the quality of their final translation for each setup [Questions F.14-15, F.22-23, F.32-33]. Task satisfaction results in Table 2 indicate that PE was the most satisfactory setup, followed by SCRATCH and STYLE in final place. Quality satisfaction results (Table 2) indicate a similar trend, with PE providing the highest quality satisfaction, followed by STYLE and then SCRATCH. These results seem to be in direct opposition to those discussed in the previous paragraph, where the majority of participants preferred the SCRATCH setup, followed by STYLE and PE. It should be noted, however, that the means are always between 1 and 3, indicating participants were not dissatisfied with any of the setups or resulting quality.

We hypothesise that the result disparity just described is due to participants linking satisfaction more strongly to the texts that were translated rather than the setups they were translated in. Table 3 shows the results of braking the satisfaction down by text. GARFIELD

SETUP	MEAN RATING	σ
SCRATCH	2.2	0.4
PE	1.9	0.6
STYLE	2.4	0.9

(a) Task satisfaction

SETUP	MEAN RATING	σ
SCRATCH	2.7	0.9
PE	1.9	0.8
STYLE	2.4	1.2

(b) Quality satisfaction

Table 2: Mean ratings and standard satisfaction ratings with the task and the final translation quality broken down by setup. Ratings range from 1 (Very satisfied) to 5 (Very dissatisfied).

TEXT	MEAN RATING	σ
AI	2.4	0.7
CHARLOTTE	2.4	0.7
GARFIELD	1.9	0.6

(a) Task satisfaction

TEXT	MEAN RATING	σ
AI	2.7	0.9
CHARLOTTE	2.4	1.0
GARFIELD	1.9	1.1

(b) Quality satisfaction

Table 3: Mean ratings and standard deviation for satisfaction ratings with the task and the final translation quality broken down by text. Ratings range from 1 (Very satisfied) to 5 (Very dissatisfied).

always comes out top, which is in line with it being the most liked text (Section 9.4.2). A larger participant pool would be required to obstain more conclusive results.

9.4.2 *Like/Dislike, Easy/Difficult and Control*

To provide more context to the satisfaction ratings, participants were asked what they liked or disliked about the setups [Questions F.12-13, F.18-19, F.26-27]. They were also asked what translations where the hardest and the easiest [Questions F.3-6] and during which setups they felt they had the most and least control [Questions F.7-10]. We discuss each of these separately.

The majority of participants mentioned textual considerations as a factor in deciding whether a text was easy or hard [Questions F.3-6]. These included text structure, topic, terminology, etc. When asked why a text/setup was hard, one participant complained that editing the MT suggestion doubled the amount of work compared to translating from scratch. When asked why they had chosen a text/setup as the easiest, two participants mentioned that post-editing an MT suggestion was faster and easier. When braking down the results by text, GARFIELD is mentioned as the easiest (56 % of participants) and is the chosen the least times as the most difficult (only one participant).

CHARLOTTE turns out to be the most difficult if we combine that it's in joint first place for the most difficult (AI and CHARLOTTE have 4 counts each) and it receives just a single count as the easiest.

The case of the AI is interesting. Considered both the hardest (4 participants) and the second easiest (3 participants) to translate, the data shows no correlation between its hard/easy consideration and having translated it in a particular setup. This further strengthens the view that difficulty is linked to a certain text rather than a certain setup, although as previously mentioned setups can play a part in this.

As for what was liked or disliked, participants mentioned more frequently in their open answers aspects relating to the text topic or specific textual elements. Specifically, GARFIELD seems to have been the text most liked by participants. The quality and usefulness of the MT suggestions also appear fairly frequently. Once again, it seems that the texts have more of an influence on what participants liked or disliked than specific.

Finally, the data on control or lack of it paints a similar picture. When asked where they felt they had the most control [Question F.7-8], responses were spread equally among setups but showed great variation in texts: GARFIELD made them feel most in control, followed by AI and CHARLOTTE. Open answers present as reasons a mix of textual considerations (participants knew a lot about Garfield, there were no tricky terms) and setup considerations (some disliked have suggestions or hints of any kind). As for where they had least control [Questions F.9-10], results again show a flat variation among setups but differences in texts: AI made them feel the least in control, followed by CHARLOTTE and GARFIELD. Here, the open answers overwhelmingly mention textual considerations such as the topic and terminology.

9.4.3 *Causes*

To throw further light onto the high acceptance of the PE setup, we can look at the quality ratings for the MT suggestions. Following Carl, Gutermuth, and Hansen-Schirra (2015) and what is commonplace in MT evaluation, participants were asked to rate the MT quality on three criteria: grammaticality, style and accuracy [Question F.20]. Results (Table 4) show all three criteria were considered average, edging towards above average. There were even some counts of well above average grammaticality (1 count) and style (1 count). This in contrast to participants in Carl, Gutermuth, and Hansen-Schirra (2015), who rated all criteria closer to below or well below average. A perceived higher quality of the MT suggestions explain the higher acceptance of PE in this study. As for the STYLE setup, further discussion is provided in Section 9.5.

CRITERIA	MEAN RATING	σ
Grammaticality	2.7	1.0
Style	2.6	0.7
Accuracy	3.0	1.0

Table 4: Mean ratings and standard deviation for MT quality evalutaions. Ratings ranged from 1 (Well above average) to 5 (Well below average), with 3 being Average.

9.4.4 *Summary of Preferred Setups*

Thus, after the previous considerations, the data indicated that participants' preferred setup was SCRATCH, followed by STYLE and PE. This is essentially based on Questions F.21 and F.31.

The previous findings also make it is reasonable to suppose that participants perceive the specific text (topic and textual considerations) as having the most influence in their preferences and enjoyment of a translation (degree of easiness, like/dislike, and task and quality satisfaction). This is useful insight for designing questionnaires whose aim is to evaluate a tool. Although the free form answers do provide some insight into the tool, it is best evaluated with a direct question of the kind "'Whould you have preferred to translate without *X*?".

9.5 STYLECHECK EFFECTIVENESS

This section evaluates the efectiveness of StyleCheck by analysing the actual translations produced by the participants. Before delving into the details, it is important to determine whether participants considered that sticking to the Wikipedia SG rules was important (just as the translation brief stated it was).

9.5.1 *Style Guide Importance*

Participants were asked if they had established any priorities and restrictions after reading the brief and, if yes, which ones [Questions F.1 and F.2]. The notion of priorities and restrictions is taken from Zabalbeascoa (1999), and describes the hierarchical list of aims and goals that translators establish for a translation after taking all factors into account (ST, target culture, client requirements, deadline and even pay). 67% reported having established priorities and restrictions. Out of these, half explicitly mentioned following the Wikipedia Style Guide as their top prority. One participant mentioned specific aspects of the writing style used in Wikipedia, and the rest mentioned gen-

eral strategies of translation. Thus, the translation brief succeeded in making most participants prioritise style guide application. The others may have not prioritised it considering that the translations were short and they were not receiving compensation for the experiment, making it not worth the time and effort to read the whole style guide and follow it. This would also fit in with the model of priorities and restrictions, which as mentioned before allows for time and payment considerations to come into play.

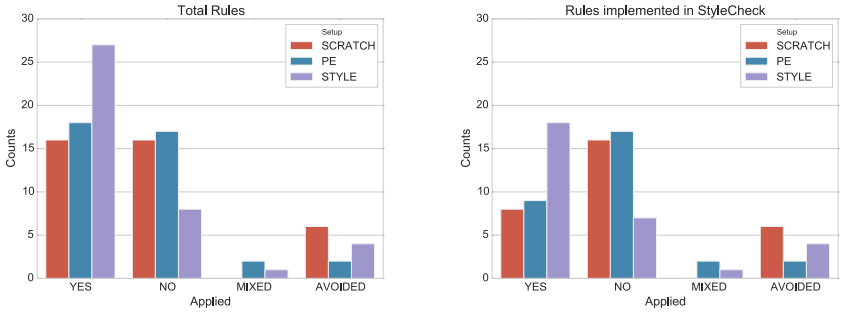
Participants were asked if they had consulted the Wikipedia style guide linked in the brief [Question F.28]: 67 % said they had. Logging carried out on the server indicates the link to the Wikipedia Style Guide was only clicked by a third of the participants, one of which answered that they hadn't consulted the style guide. This could be due to participants opening up a new tab and searching for the style guide themselves, but given that the link was provided on all translation pages, this seems unlikely. It seems participants felt they should have consulted it, but preferred to say they had when they actually had not. This is a strong indicator of the usefulness of a digitally applied style guide: participants felt compelled to apply it, but did not. As stated before, this was probably due to the lack of time and short texts which did not make up for the investment in time required to read and apply the style guide.

When asked why [Question F.29], those who hadn't consulted the guide said they were not used to using style guides or that they had previously translated Wikipedia articles and were familiar with it. Those who had, either stated the general importance of using a style guide when one is provided by a client or mentioned specific elements they looked up (use of italics, parenthesis, how to handle names of works of art, etc.).

9.5.2 *Style Guide Application*

To evaluate how well StyleCheck worked, we first look at the objective data of whether a particular rule was applied or not in each translation. Figure 13a presents how many times the rules identified in the texts were applied or not. The MIXED category refers to elements that appear more than once in a single text and present inconsistencies: some instances follow the associated style rule, others don't. The AVOIDED category refers to instances where the text was changed, resulting in the rule issue being avoided. As can be seen, the STYLE setup leads to a notably higher amount of rules applied than the other setups. This is further confirmed in Figure 13b, which shows the data only for rules that were implemented in StyleCheck.

Rule application varied a lot depending on the particular rule, as can be seen in Figure 14. Some rules, such as 3B and 3C were always applied. They were related to aspects (number formatting) that



(a) Total rules (b) Rules implemented in StyleCheck

Figure 13: Number of rules that were applied in the final translations. Both the total number of rules detected in the texts and the subset of rules implemented in StyleCheck are presented.

translators usually know are contained in style guides and may have previous experience in how to handle. Others, such as 2A and 2B were generally not applied. These were related to aspects very specific to the Wikipedia style guide (chronologiclly ordering lists and avoiding time expressions that refer to the moment of enunciation).

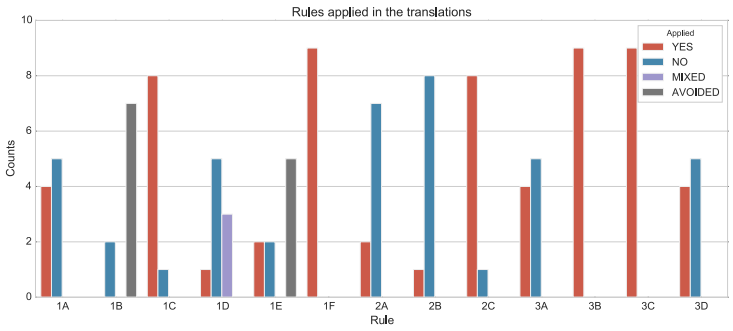
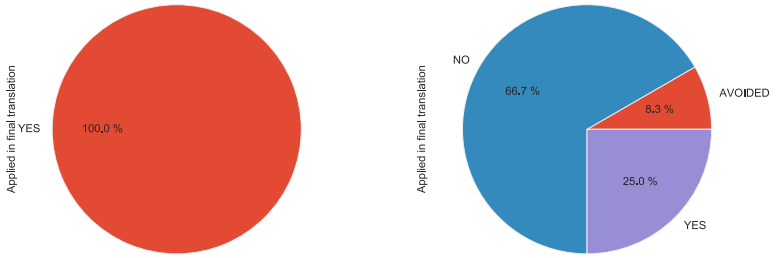


Figure 14: Number of rules applied in the final translations for all setups. Counts are broken down by rule.

9.5.3 *StyleCheck vs. PE*

We now turn to a comparison of the STYLE and PE setups. Starting off with PE, we find that participants tended to closely follow what the MT suggestion presented. Figure 15a shows that all instances of rules that were applied in the MT suggestion were also applied in the final translation. Figure 15b shows similar results: 67 % of rules not applied in the suggestion were not applied in the final translation. Only 25 % of them were corrected so that they were correctly applied. The mixed instances in the MT suggestion provide further evidence: two out of the three instances were also carried on to the final trans-

lation. With a simple revision of the final translation, participants should have noticed the inconsistencies and homogenised them.



(a) Rules applied in MT (b) Rules not applied in MT

Figure 15: Percentage of rules applied or not in the final translation for the PE setup depending on whether theMT suggestion had applied them or not.

Delving further in the data, we find that StyleCheck was effective in getting participants to follow the suggestions. When presented with a suggestion (indicating the participant had written a sentence that did not follow a specific rule), 71 % of the time participants corrected their translations so that they followed the suggestion.

To find out why in 29 % of the cases the suggestions weren't followed, we can check two sources of data, objective and subjective. Objective data indicates that out of the five cases where the suggestion wasn't followed, in two cases the relevant suggestion was shown in third position below two other suggestions. It is possible that depending on the screen resolution and browser window configuration the suggestion appeared outside of the screen and was not seen or ignored. The fact that some suggestions were long could have also contributed to this. A further two cases presented the suggestion in first position, indicating the participant ignored the suggestion or chose not to follow it.

Subjective data was collected through the final questionnaire [Question F.30] and is presented in Table 5. Participants rated their agreement with various satements from 1 to 5. Participants generally agreed that the style hints were helpful, and agreed that the information they contained was right. Results show participants generally thought the style hints were easy to understand, but tended to agree that they were too long. This would suggest the hints could be improved by rewriting them into a tenser alert style, rather than copy the original style guide description. Regarding the user interface, participants on average remained rather neutral on whether there were too many hints and whether the boxes were distracting. However,

both considerations included one instance of a participant strongly agreeing.

CONSIDERATION	MEAN AGREEMENT	σ
The style hints were helpful	2.3	1.0
There were too many style hints	2.8	1.1
The style hints were easy to understand	1.6	0.7
The style hints were too long	2.1	1.1
The boxes where the style hints were shown were distracting	2.6	1.1
The style hints contained wrong information	3.7	1.0

Table 5: Mean agreements and standard deviations for whether participants agree or not with various statements related to the StyleCheck suggestions. Ratings range from 1 (Strongly Agree) to 5 (Strongly Disagree).

9.6 FUTURE WORK

9.6.1 *Two Types of PE*

The data revealed that PE seems to only improve translation speed when little editing is done, what has been called light PE. Full PE, on the other hand, doesn't necessarily amount to a large speed improvement and forces a workflow on translators that some dislike in comparison to translating from scratch. Thus, I would classify light PE under human-assisted MT, where the main structure and choices in the translation are selected by the system. Full PE would fall into the CAT category, but more research is required into this variant to see if the slow-downs and small speed improvements also appear in larger studies. If they do, it could question the usefulness of PE for full-quality translation.

9.6.2 *TPR Methodology*

Within the space of verbal report data, there is a need for standardised set of questions to make comparisons across studies possible. The CRITT TPR-DB (Carl, 2012) database already does this for data such as keylogging, time, eye-tracking and other measurements, but verbal-report data in the form of questionnaires, for example, is completely left out. As this thesis has shown, this information is key in providing context to the data collected from logging and tracking of all kinds.

Future studies should work in the direction of trying to use similar questions and analysing the data in a similar fashion.

9.6.3 *StyleCheck: Improvements*

The questionnaires brought up a few issues with the suggestion interface StyleCheck uses. Notably, it would be better for the text to be shorter and more to the point. Future versions of StyleCheck should rewrite the raw text included in a style guide to better adapt to the needs of a quick suggestion, which could include a link to the full description if needed. Another area of improvement is with regards to the always-on suggestions. Some participants stated they would prefer to completely dismiss a suggestion or all of them, so this functionality should be included. Optionality of StyleCheck itself is also important given that a number of participants stated they would have preferred to translate without them.

9.6.4 *StyleCheck: MT Evaluation Metric and Beyond*

The GF-based approach to implementing style guides can be useful for other tasks. The most interesting and useful is as a metric for MT evaluation. StyleCheck can be turned into a metric which operates in a similar fashion to unit testing in the software development world.

Style guides, as part of a translation brief, encode requirements and expectations the translation has to fulfill. These requirements are embodied in isolated GF functions that can be individually checked to see whether a rule was applied or not. GF is flexible enough to allow many different kinds of rules to be created and check whether many linguistic aspects appear in a text or not. An MT metric can be built upon this basis, checking each translation for all the aspects relevant to a particular client, domain, text type, etc. Thus, the notion of "quality" becomes flexible and adaptable according to translator, client and situational needs, more in-line with the skopos paradigm in translation theory.

This metric would have numerous advantages. It would allow to see what specific aspects a translation failed in. Through attaching a weight to each aspect, a hierarchy can be established to prioritise particular aspects over others, in a similar way to the model of priorities and restrictions (Zabalbeascoa, 1999). Once the metric is defined, it can be used to rank candidates output by a decoder in an SMT system or translation options from multiple systems, as well as being used in SMT tuning.

Lastly, StyleCheck can also be used for automatic post-editing of certain aspects that don't require translator intervention, as explained in Chapter 6.

10 CONCLUSION

This thesis has presented a unified methodology for developing and evaluating CAT tools. The methodology has then been used to develop StyleCheck, a tool that helps translators by giving them hints about style guide rules that should be applied while they are translating.

10.1 STYLECHECK

Results show StyleCheck acheived its main goal: leading to a higher rate of rule application. Using it lead to more style rules being applied when compared with translation from scratch and post-editing.

StyleCheck does not seem to burden the translation process. Translators prefer it to post-editing. Speed is comparable to from-scratch translation, but the UI needs some work. Although not considered distracting nor intrusive, the interaction can be improved upon, especially with regards to adapting and rewriting the style guide rules to make them terser.

Despite the advantages, developing the rules themselves requires a lot of manual work. Inventoring all the options that a rule should match can be tricky, and some instances can fall through the net.

10.2 METHODOLOGY

The CAT tool development methodology described in this thesis proved its worth. A simple survey among translators came up with plenty of suggestions for CAT tool improvements, as did a quick read through translation theory.

Methods developed in TPR, when used in combination (triangulation), managed to generate a wealth of data into many aspects of a tool: how translators use it, how it affects the way they translate and the impact on the final translation. This was possible even considering the only the most basic TPR methods were used; more advanced methods such as eyetracking and keylogging should further improve understanding of how a tool is used.

As for specific findings related to evaluation, it has been shown that in questionnaires it is best to avoid questions related to ease/difficulty or like/dislike, as the text itself can weigh more heavily than the setup used. It is thus recommended to use questions that directly ask if translators would prefer to translate without a certain tool.

Continuing on the topic of evaluation, time as a metric should be used with care. Other studies that claim speed improvements in

some setups or using certain tools fail to consider translation as a whole and ignore time-consuming tasks. Data presented in this thesis showed that post-editing speed increases only apply to the light variant with minimal changes. Full post-editing, where it is used as a CAT tool, doesn't show large speed-ups and can even slow down some translators.

Prototyping as used in this thesis to develop StyleCheck is shown to be very useful as a feedback loop. A relatively modest investment in time and effort to build a prototype still manages to generate sufficient data about how the full tool will perform. Prototyping also allowed for problems with the tool (such as the hints being too long and always-on) to surface at an early stage so they can be quickly fixed.

All in all, work carried out in this thesis provides solid ground for the CAT tool development methodology to be adopted and used in future research.

Part IV

APPENDIX

A

EXPERIMENT WALKTHROUGH

The following are screenshots of each page of the experiment website. First, participants were greeted and given an overview of the experiment procedure (Figure 16). Next, participants were asked to complete the initial questionnaire (Figure 17). After submitting it, they were presented with the translation brief for the three translation tasks on a page on its own (Figure 18). The aim was for participants to read it and pay attention to it before starting the translations. Participants were told that the brief would be available on the following pages should they need to read it again.

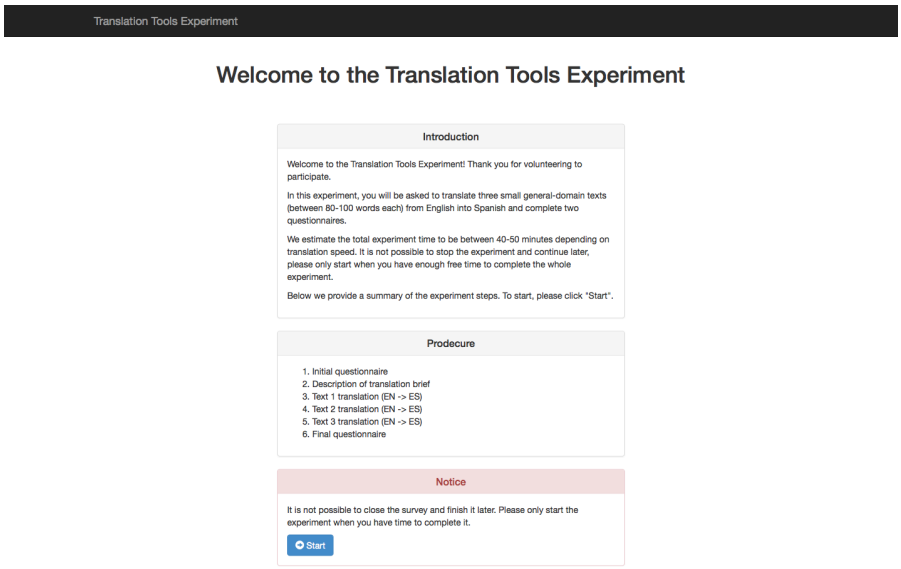


Figure 16: Website (Page 1). Main page with a general overview of the experiment.

Then, participants had to carry out the three translations. At the top of each page, instructions were provided for each setup and the translation brief was included. First, the SCRATCH setup (Figure 19). Second, the PE setup (Figure 20). Third, the STYLE setup (Figure 21). The texts were presented according to the order participants had been provided in the initial link.

Finally, participants were asked to complete the final questionnaire (Figure 22). They were provided with the ST of the translations they had completed (and the MT suggestion in the PE setup) to help jog their memory. Finally, participants are thanked for their participation (Figure 23).

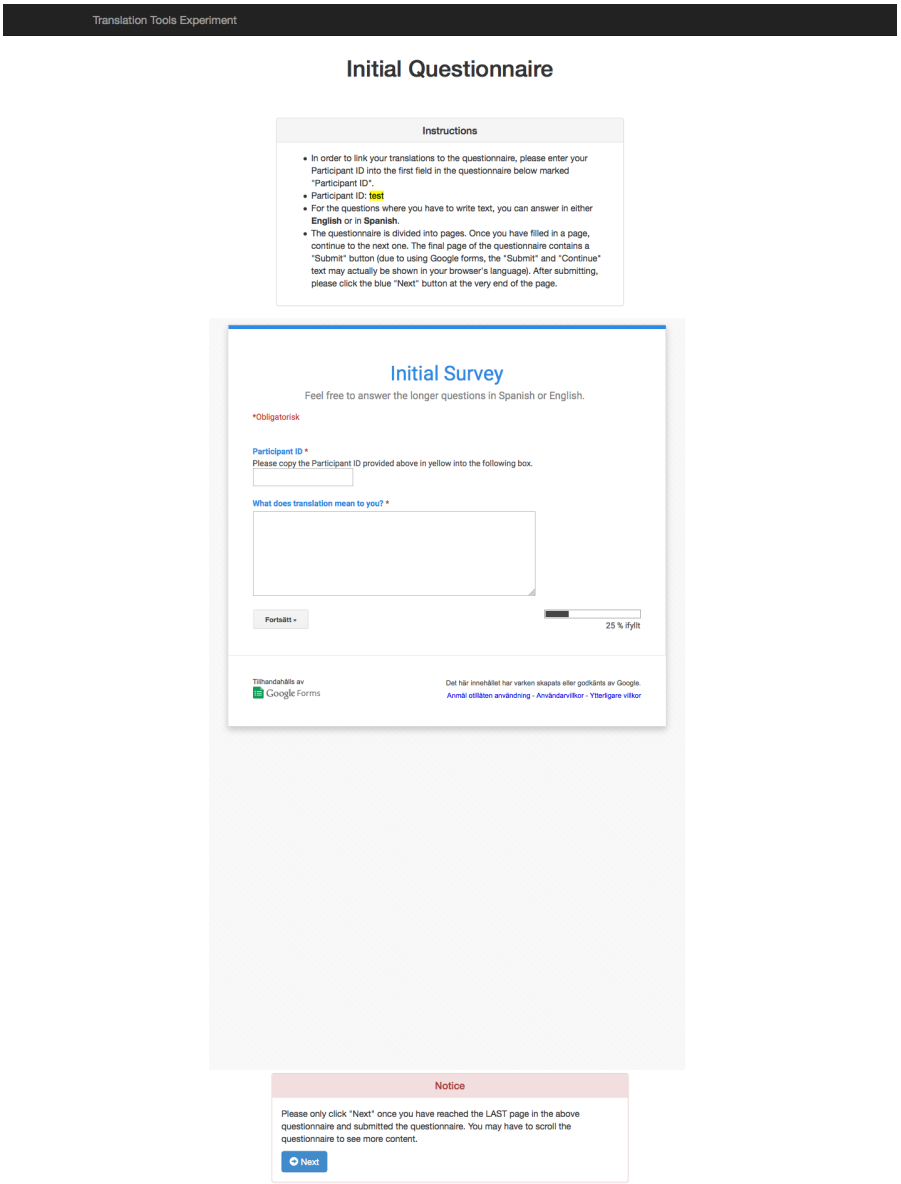


Figure 17: Website (Page 2). Participants answer the initial questionnaire.

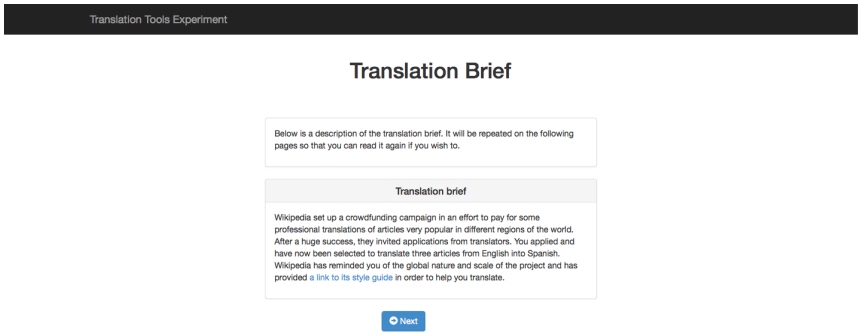


Figure 18: Website (Page 3). Translation brief description.

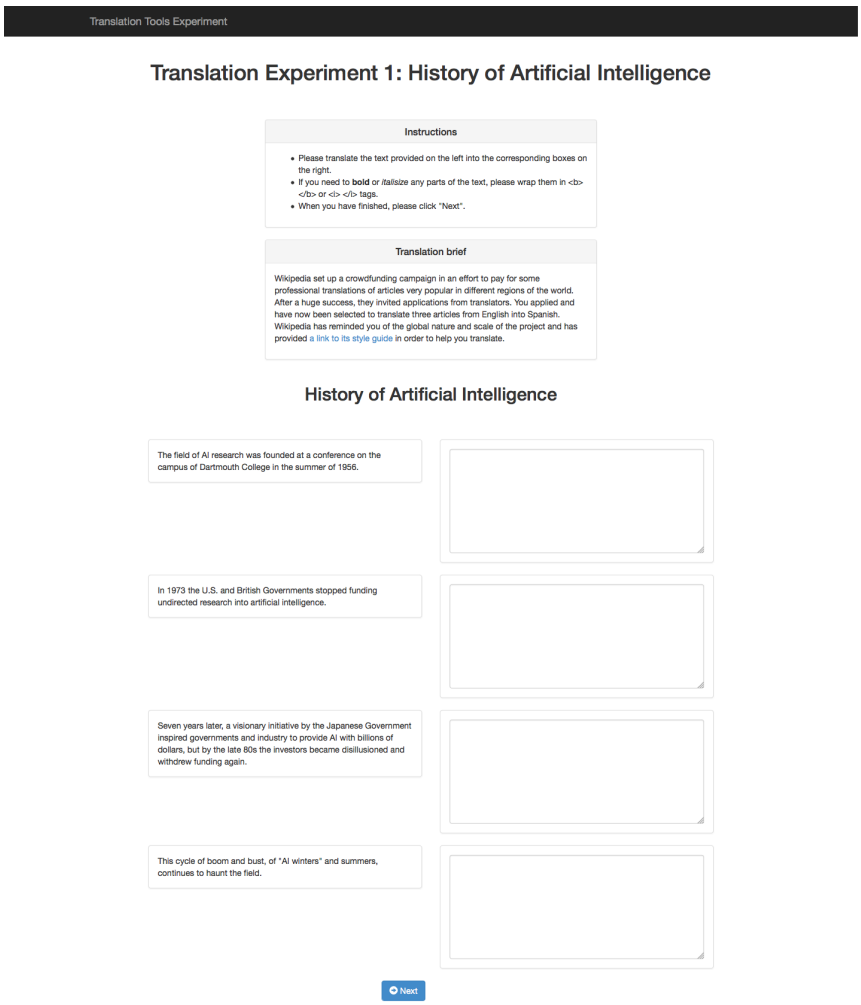


Figure 19: Website (Page 4). Participants translate in the from-scratch setup.

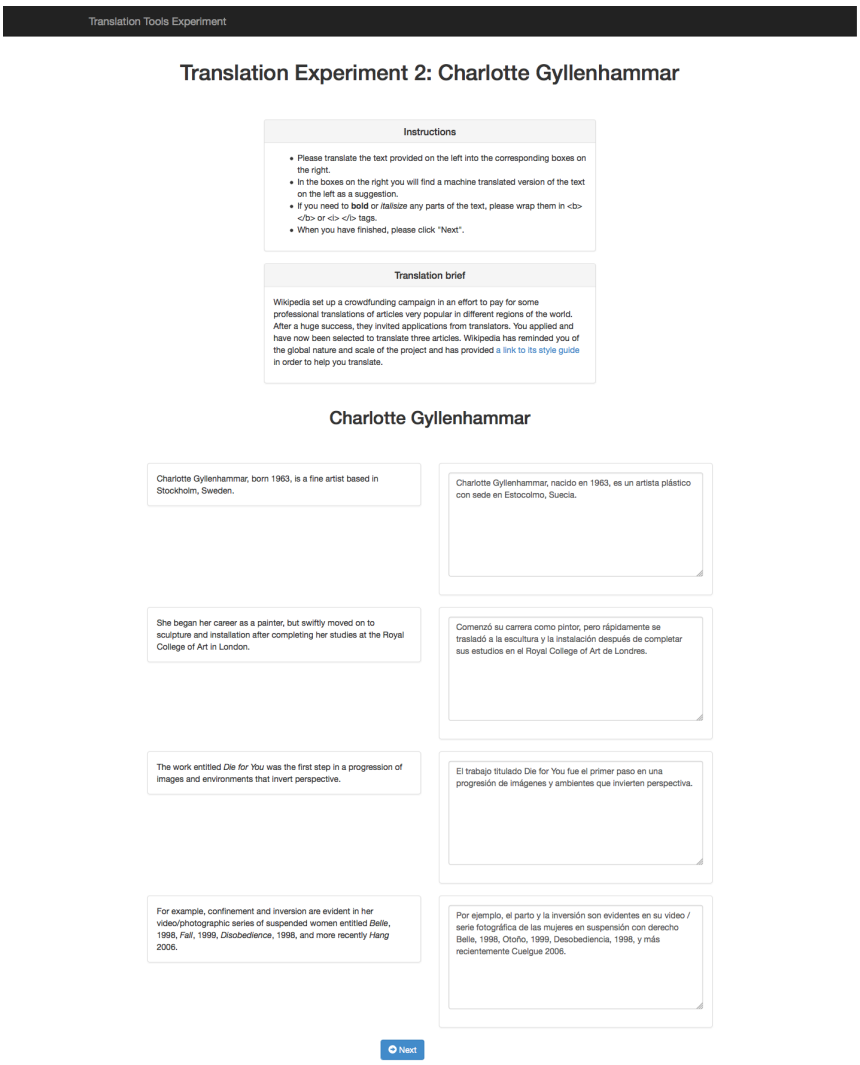


Figure 20: Website (Page 5). Participants translate in the PE setup.

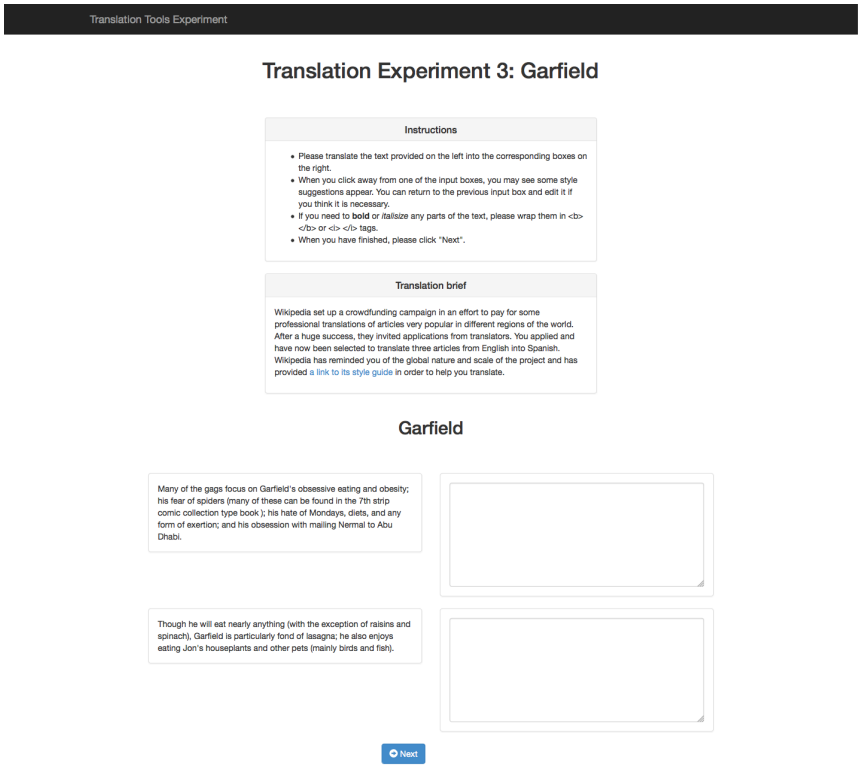


Figure 21: Website (Page 6). Participants translate in the style setup.

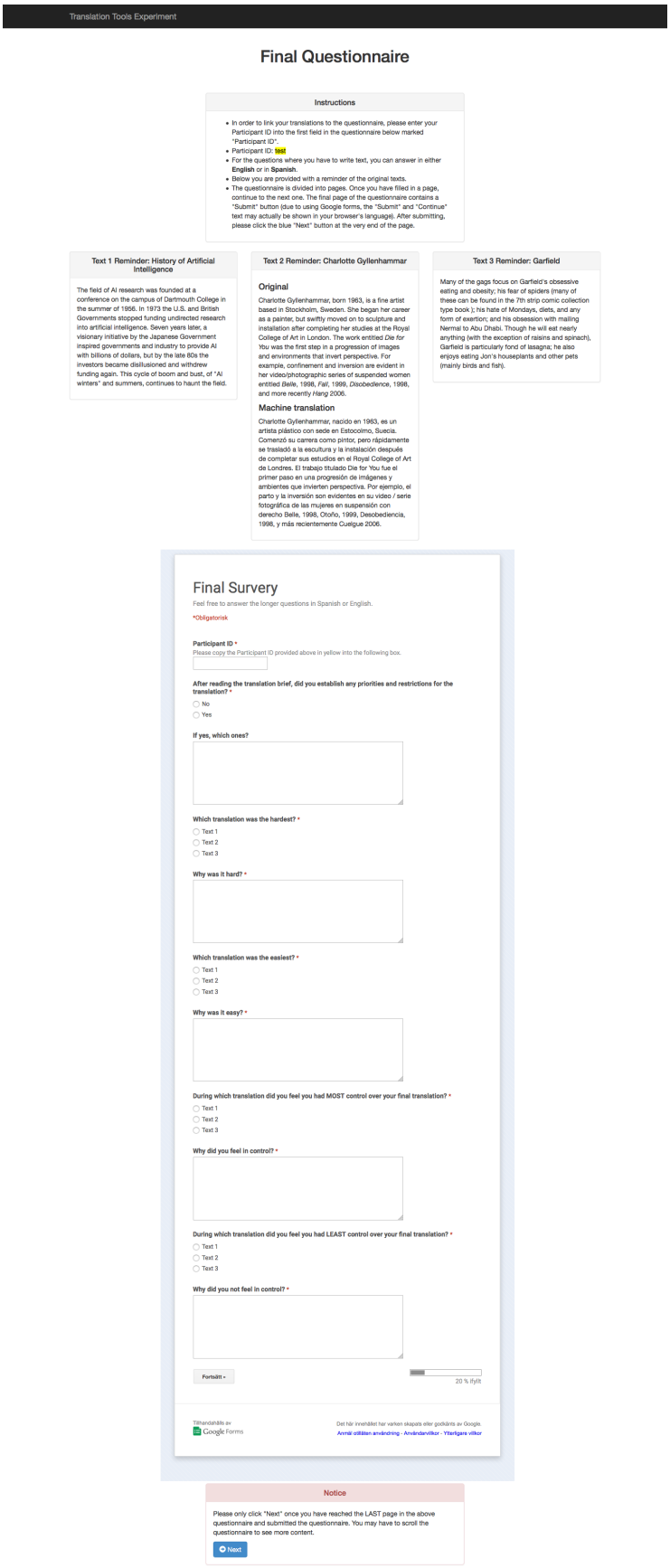


Figure 22: Website (Page 7). Participants are asked to answer the final questionnaire.

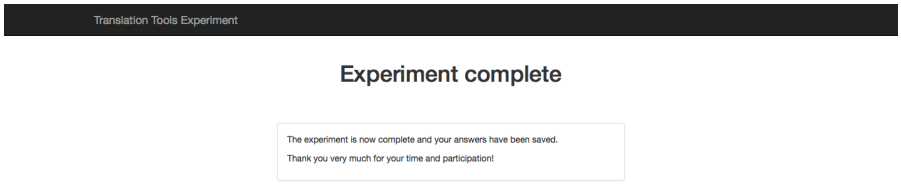


Figure 23: Website (Page 7). Participants are thanked for their collaboration.

B

QUESTIONNAIRES

This appendix contains screenshots of both the initial (before translating) and final (after translating) questionnaires that were given to the study's participants. For further details on the design of the questionnaires, please refer to Chapter 8.

PLEASE NOTE: Form navigation and interface elements, such as the "Next" and "Submit" buttons and "Required*" field, were shown in the language participants have their Google account set up in or, if they don't have one, in the browser's default language. The screenshots in this chapter show the navigation and interface elements in Swedish, but this varied for each participant.

B.1 INITIAL QUESTIONNAIRE

B.1.1 *Questions*

Note: in case of closed questions, the possible options are shown in square brackets.

1. What does translation mean to you?
2. What is your gender? [Male, Female, Other]
3. What is your year of birth?
4. What is your current occupation?
5. Which language(s) are you native in? [English, Spanish, French, German, Catalan, Basque, Galician, Chinese, Japanese, Portuguese, Italian, Russian, Greek, Other]
6. Which other language(s) do you speak? [English, Spanish, French, German, Catalan, Basque, Galician, Chinese, Japanese, Portuguese, Italian, Russian, Greek, Other]
7. Have you studied translation? [Yes, No]
8. If yes, for how many years have you studied translation?
9. Have you worked as a translator? [Yes, No]
10. If yes, for how many years have you worked as a translator?
11. Do you have professional experience working on the following translation-related tasks? [Multiple choices can be selected] [Translating from scratch, Proof-reading, Machine translation post-editing, Transcreation, Loclalizing, Managing terminology,

Setting-up style guides, Translating with a CAT tool (For example, Trados), Preparing documents for internationalisation, Machine translation system setup (for example, Moses SMT), Machine translation quality evaluation]

12. What do you think are the most INTERESTING tasks to carry out when translating?

13. What do you think are the most BORING tasks to carry out when translating?

14. If a tool could be created to automate any task you carry out when translating so you don't have to do it, which task would it be?

15. How important are the following considerations when translating? [Very important (1), (2), (3), (4), Not at all important (5)]
    a) Accurately portraying the source text
    b) Applying a style guide consistently
    c) Using terminology consistently
    d) Providing a grammatically correct translation
    e) Conforming to what the client requests in the translation brief
    f) Providing a fluent translation/idiomatic translation

16. Which of the previous considerations do you prioritise the most? [Accurately portraying the source text, Applying a style guide consistently, Using terminology consistently, Providing a grammatically correct translation, Conforming to what the client requests in the translation brief, Providing a fluent translation/idiomatic translation]

17. Are there any other considerations you usually take into account?

18. Do you usually use glossaries / terminological lists when translating? [Yes, No]

19. Do you usually use style guides when translating? [Yes, No]

20. Do you usually use computer-aided translation (CAT) when translating? [Yes, No]

21. Do you agree with the following statements? [Strongly agree (1), (2), (3), (4), Strongly Disagree (5)]
    a) Machine translation is useful for translators
    b) Machine translation is ONLY useful for the general public

c) Machine translation will one day replace human translators

d) Machine translation today provides good quality translations

e) It is difficult to understand how machine translation works

f) I would like to learn more about machine translation

22. Do you think your translations are better than those from machine translation? [Yes, No]

23. Why?

B.1.2 *Screenshots*

The following are screenshots of the questionnaire that participants had to fill in before starting the translation tasks.

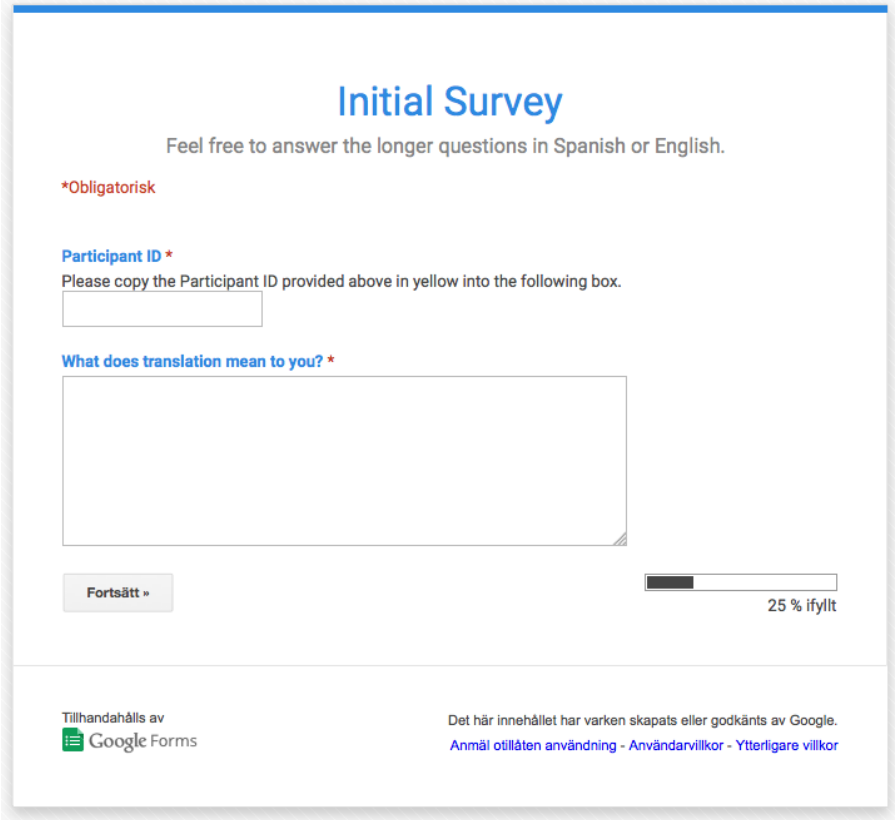
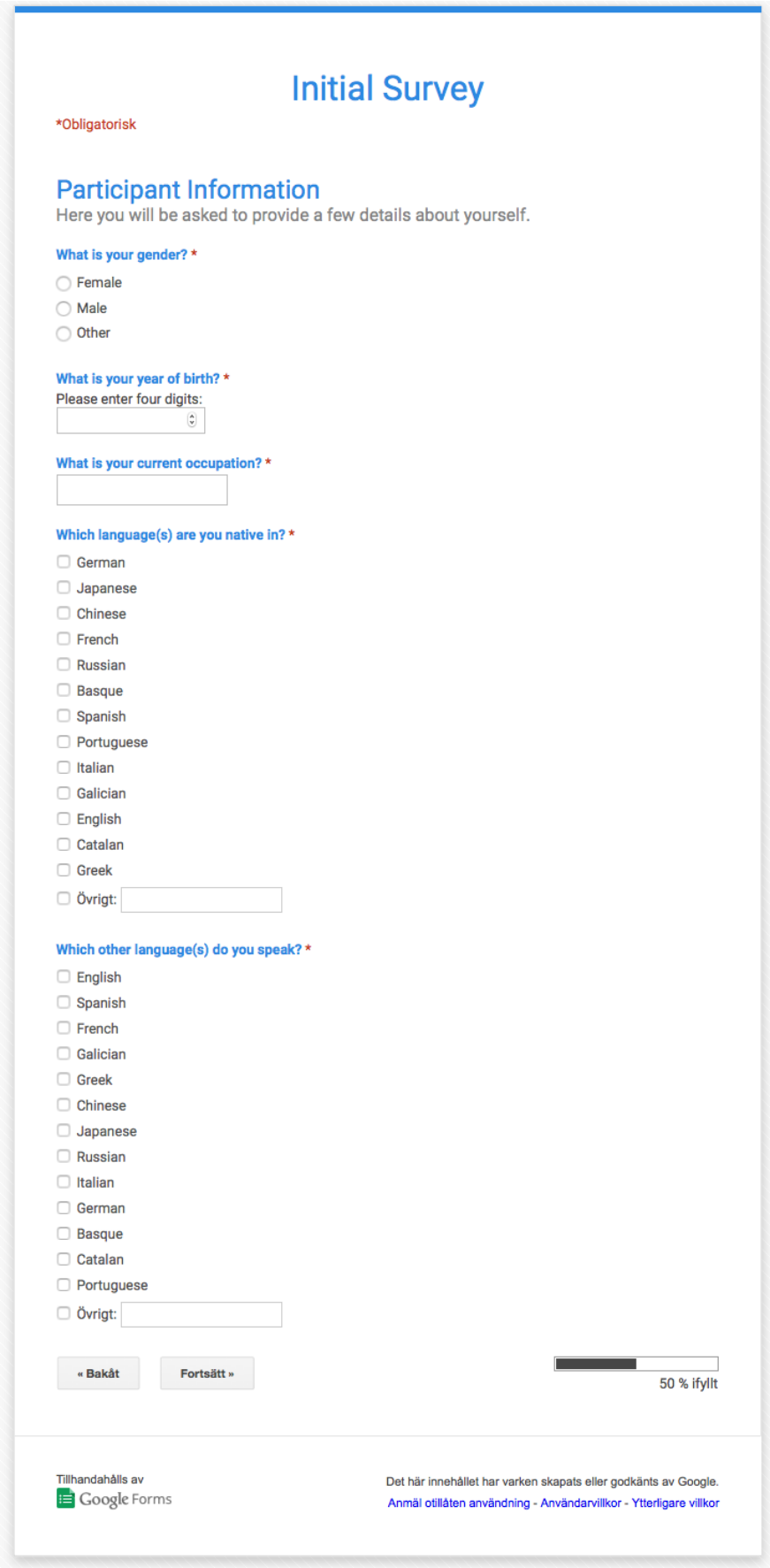


Figure 24: Initial Questionnaire (Page 1)



Initial Survey

*Obligatorisk

Participant Information

Here you will be asked to provide a few details about yourself.

What is your gender? *

- Female
- Male
- Other

What is your year of birth? *

Please enter four digits:

What is your current occupation? *

Which language(s) are you native in? *

- German
- Japanese
- Chinese
- French
- Russian
- Basque
- Spanish
- Portuguese
- Italian
- Galician
- English
- Catalan
- Greek
- Övrigt:

Which other language(s) do you speak? *

- English
- Spanish
- French
- Galician
- Greek
- Chinese
- Japanese
- Russian
- Italian
- German
- Basque
- Catalan
- Portuguese
- Övrigt:

« Bakåt Fortsätt »

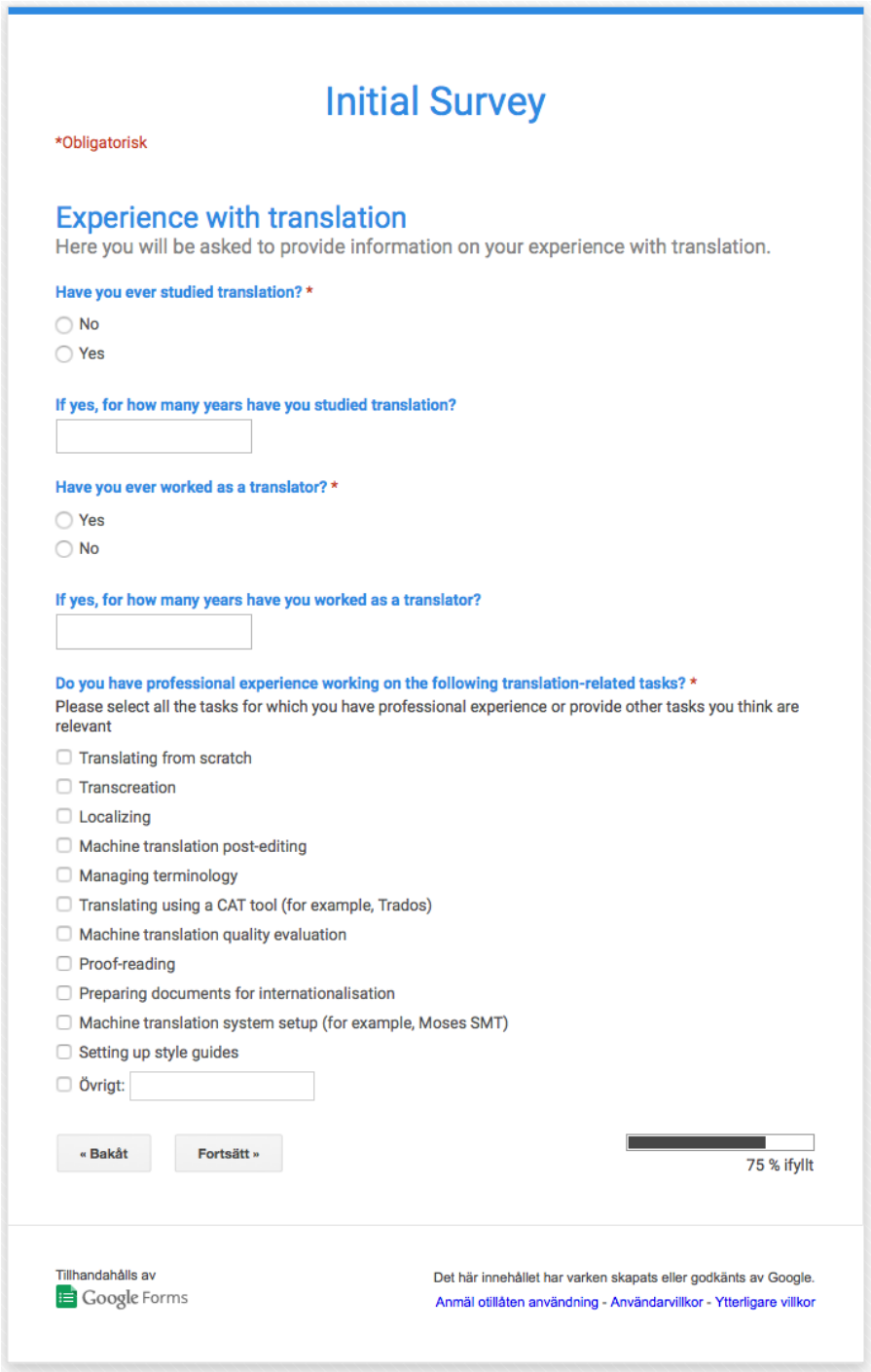
50 % ifyllt

Tillhandahålls av Google Forms

Det här innehållet har varken skapats eller godkänts av Google.

Anmäl otillåten användning - Användarvillkor - Ytterligare villkor

Figure 25: Initial Questionnaire (Page 2)



Initial Survey

*Obligatorisk

Experience with translation

Here you will be asked to provide information on your experience with translation.

Have you ever studied translation? *

- No
- Yes

If yes, for how many years have you studied translation?

Have you ever worked as a translator? *

- Yes
- No

If yes, for how many years have you worked as a translator?

Do you have professional experience working on the following translation-related tasks? *

Please select all the tasks for which you have professional experience or provide other tasks you think are relevant

- Translating from scratch
- Transcreation
- Localizing
- Machine translation post-editing
- Managing terminology
- Translating using a CAT tool (for example, Trados)
- Machine translation quality evaluation
- Proof-reading
- Preparing documents for internationalisation
- Machine translation system setup (for example, Moses SMT)
- Setting up style guides
- Övrigt:

« Bakåt Fortsätt »

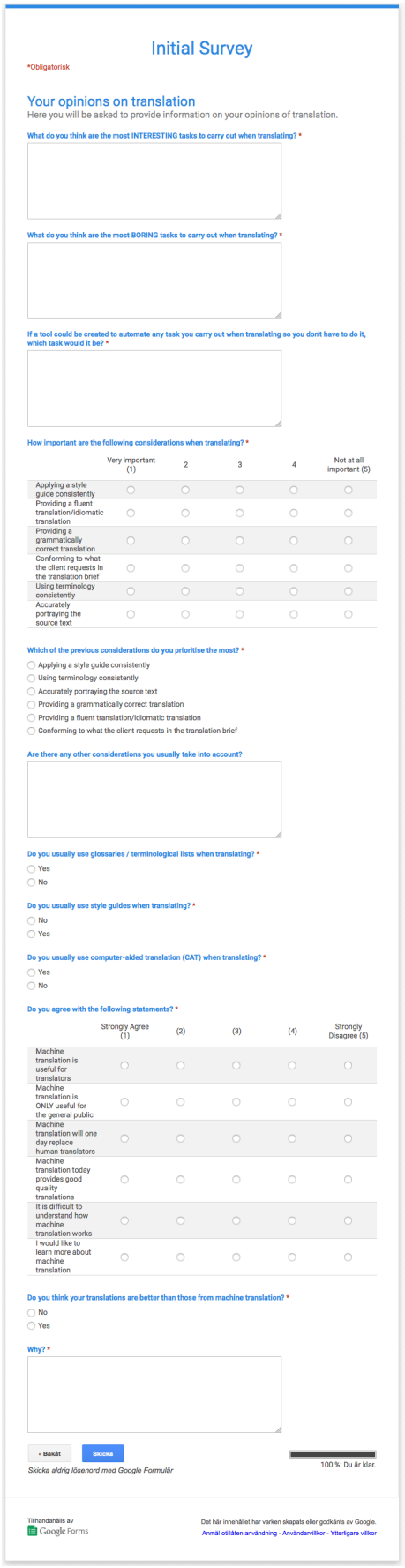
75 % ifyllt

Tillhandahålls av Google Forms

Det här innehållet har varken skapats eller godkänts av Google.

Anmäl otillåten användning - Användarvillkor - Ytterligare villkor

Figure 26: Initial Questionnaire (Page 3)



Initial Survey

*Obligatorisk

Your opinions on translation

Here you will be asked to provide information on your opinions of translation.

What do you think are the most INTERESTING tasks to carry out when translating? *

What do you think are the most BORING tasks to carry out when translating? *

If a tool could be created to automate any task you carry out when translating so you don't have to do it, which task would it be? *

How important are the following considerations when translating? *

	Very important (1)	2	3	4	Not at all important (5)
Applying a style guide consistently	○	○	○	○	○
Providing a fluent translation/idiomatic translation	○	○	○	○	○
Providing a grammatically correct translation	○	○	○	○	○
Conforming to what the client requests in the translation brief	○	○	○	○	○
Using terminology consistently	○	○	○	○	○
Accurately portraying the source text	○	○	○	○	○

Which of the previous considerations do you prioritise the most? *

- ○ Applying a style guide consistently
- ○ Using terminology consistently
- ○ Accurately portraying the source text
- ○ Providing a grammatically correct translation
- ○ Providing a fluent translation/idiomatic translation
- ○ Conforming to what the client requests in the translation brief

Are there any other considerations you usually take into account?

Do you usually use glossaries / terminological lists when translating? *

- ○ Yes
- ○ No

Do you usually use style guides when translating? *

- ○ No
- ○ Yes

Do you usually use computer-aided translation (CAT) when translating? *

- ○ Yes
- ○ No

Do you agree with the following statements? *

	Strongly Agree (1)	(2)	(3)	(4)	Strongly Disagree (5)
Machine translation is useful for translators	○	○	○	○	○
Machine translation is ONLY useful for the general public	○	○	○	○	○
Machine translation will one day replace human translators	○	○	○	○	○
Machine translation today provides good quality translations	○	○	○	○	○
It is difficult to understand how machine translation works	○	○	○	○	○
I would like to learn more about machine translation	○	○	○	○	○

Do you think your translations are better than those from machine translation? *

- ○ No
- ○ Yes

Why? *

« Bakåt Skicka

100 %: Du är klar.

Skicka aldrig lösenord med Google Formulär

Tillhandahålls av Google Forms

Det här innehållet har varken skapats eller godkänts av Google.

Anmäl otillåten användning - Användarvillkor - Ytterligare villkor

Figure 27: Initial Questionnaire (Page 3)

B.2 FINAL QUESTIONNAIRE

B.2.1 *Questions*

Note: in case of closed questions, the possible options are shown in square brackets.

1. After reading the translation brief, did you establish any priorities and restrictions for the translation? [Yes, No]
2. If yes, which ones?
3. Which translation was the hardest? [Text 1, Text 2, Text 3]
4. Why was it hard?
5. Which translation was the easiest? [Text 1, Text 2, Text 3]
6. Why was it easy?
7. During which translation did you feel you had MOST control over your final translation? [Text 1, Text 2, Text 3]
8. Why did you feel in control?
9. During which translation did you feel you had LEAST control over your final translation? [Text 1, Text 2, Text 3]
10. Why did you not feel in control?
11. Approximately, how many minutes do you think it took you to translate Text 1?
12. What did you LIKE the most about translating Text 1?
13. What did you DISLIKE the most about translating Text 1?
14. Overall, how satisfied were you with the task of translating Text 1? [Very satisfied, Somewhat satisfied, Neutral, Somewhat dissatisfied, Very dissatisfied]
15. Overall, how satisfied were you with the quality of your final translation of Text 1? [Very satisfied, Somewhat satisfied, Neutral, Somewhat dissatisfied, Very dissatisfied]
16. Do you have any additional comments on the translation of Text 1?
17. Approximately, how many minutes do you think it took you to translate Text 2?
18. What did you LIKE the most about translating Text 2?
19. What did you DISLIKE the most about translating Text 2?
20. How would you rate the quality of the machine translation suggestions? [Well below average, Below average, Average, Above average, Well above average]

   a) Grammaticality
   b) Style
   c) Accuracy

21. Would you have preferred to translate from scratch without the machine translation suggestions? [Yes, No]

22. Overall, how satisfied were you with the task of translating Text 2? [Very satisfied, Somewhat satisfied, Neutral, Somewhat dissatisfied, Very dissatisfied]

23. Overall, how satisfied were you with the quality of your final translation of Text 2? [Very satisfied, Somewhat satisfied, Neutral, Somewhat dissatisfied, Very dissatisfied]

24. Do you have any additional comments on the translation of Text 2?

25. Approximately, how many minutes do you think it took you to translate Text 3?

26. What did you LIKE the most about translating Text 3?

27. What did you DISLIKE the most about translating Text 3?

28. Did you consult the Wikipedia Style Guide linked to in the Translation Brief? [Yes, No]

29. Why?

30. Do you agree with the following statements? [Strongly agree (1), (2), (3), (4), Strongly Disagree (5)]
   a) The style hints were helpful
   b) There were too many style hints
   c) The style hints were easy to understand
   d) The style hints were too long
   e) The boxes where the style hints were shown were distracting
   f) The style hints contained wrong information

31. Would you have preferred to translate from scratch without the style suggestions? [Yes, No]

32. Overall, how satisfied were you with the task of translating Text 3? [Very satisfied, Somewhat satisfied, Neutral, Somewhat dissatisfied, Very dissatisfied]

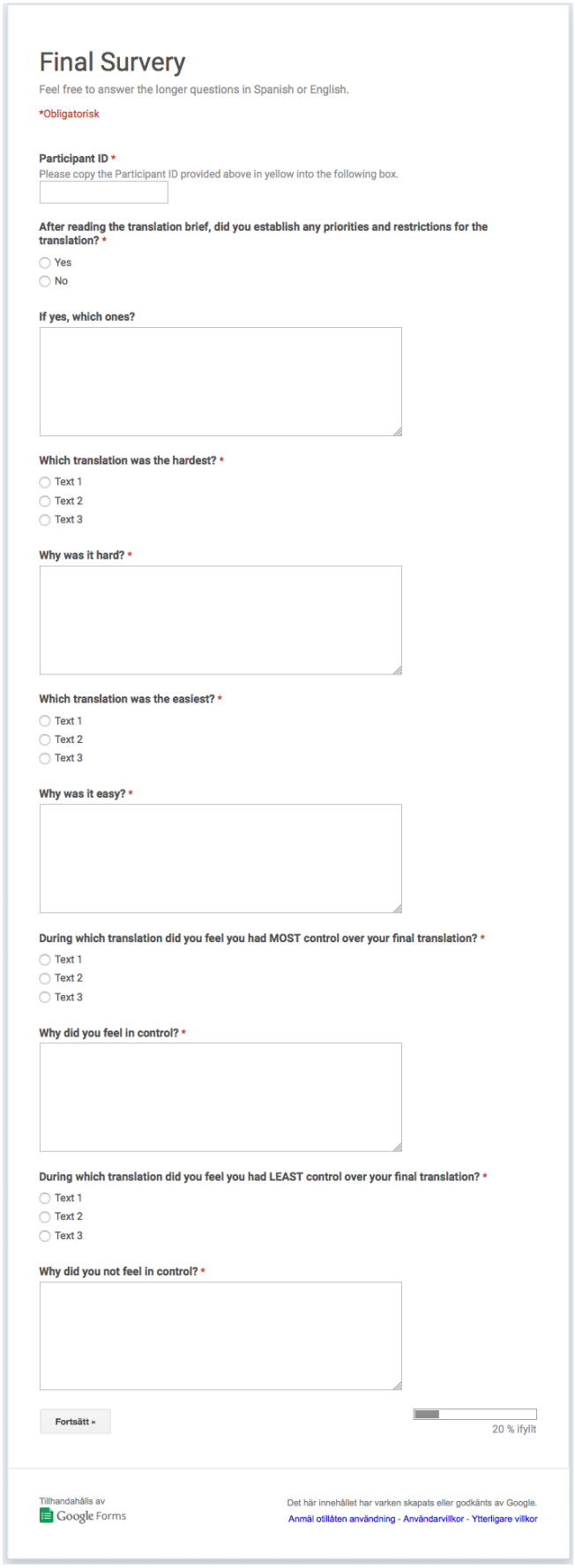
33. Overall, how satisfied were you with the quality of your final translation of Text 3? [Very satisfied, Somewhat satisfied, Neutral, Somewhat dissatisfied, Very dissatisfied]

34. Do you have any additional comments on the translation of Text 3?

35. Do you have any final comments about any aspect of the experiment?

B.2.2 *Screenshots*

The following are screenshots of the questionnaire that participants had to fill in after completing the translation tasks. Participants were shown a reminder (not seen in the following screeenshots) of the source texts they had translated to help job their memory.



Final Survery

Feel free to answer the longer questions in Spanish or English.

*Obligatorisk

Participant ID *
Please copy the Participant ID provided above in yellow into the following box.

After reading the translation brief, did you establish any priorities and restrictions for the translation? *
- Yes
- No

If yes, which ones?

Which translation was the hardest? *
- Text 1
- Text 2
- Text 3

Why was it hard? *

Which translation was the easiest? *
- Text 1
- Text 2
- Text 3

Why was it easy? *

During which translation did you feel you had MOST control over your final translation? *
- Text 1
- Text 2
- Text 3

Why did you feel in control? *

During which translation did you feel you had LEAST control over your final translation? *
- Text 1
- Text 2
- Text 3

Why did you not feel in control? *

Fortsätt »

20 % ifyllt

Tillhandahålls av Google Forms

Det här innehållet har varken skapats eller godkänts av Google.
Anmäl otillåten användning - Användarvillkor - Ytterligare villkor

Figure 28: Final Questionnaire (Page 1)

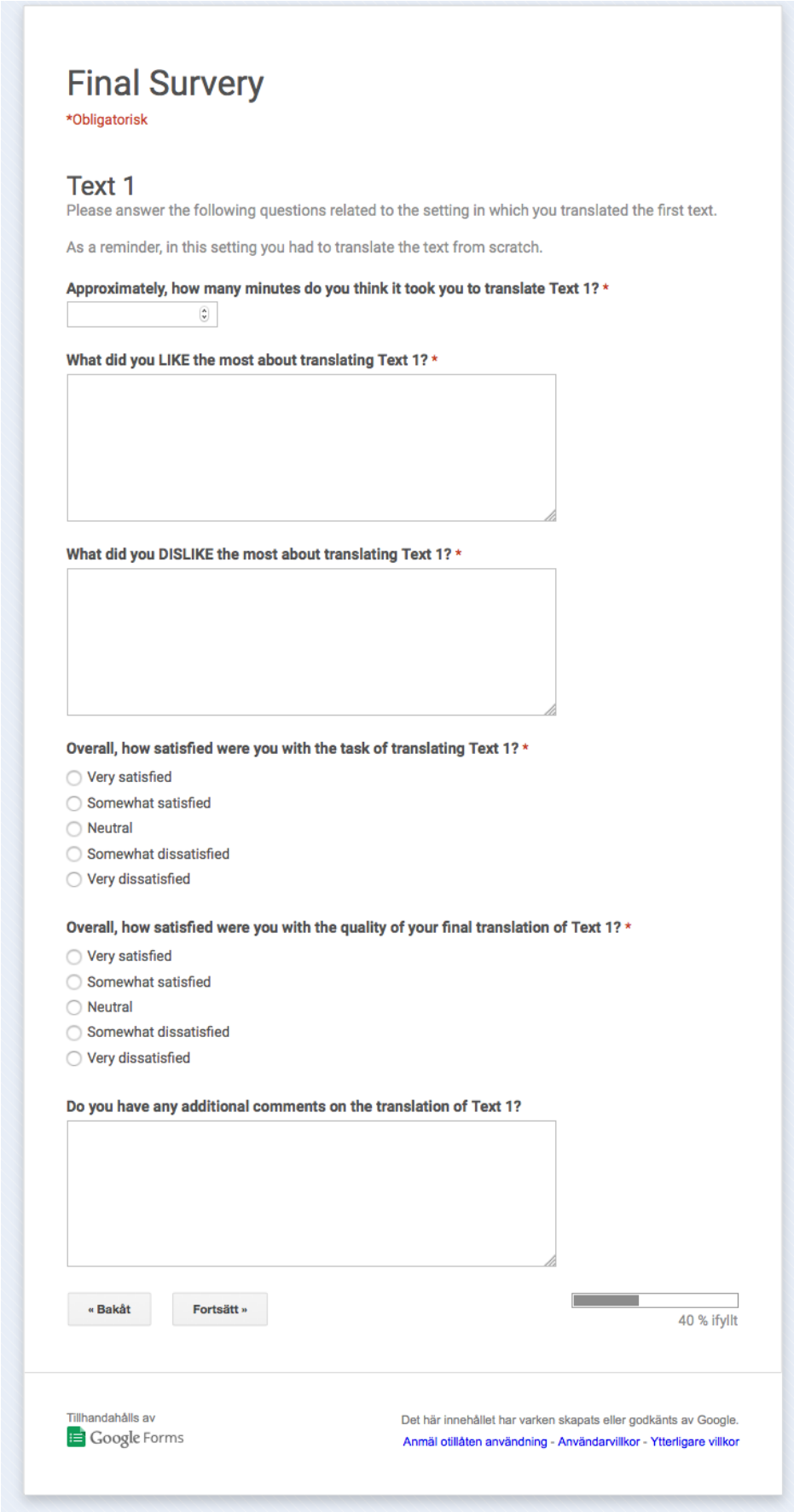
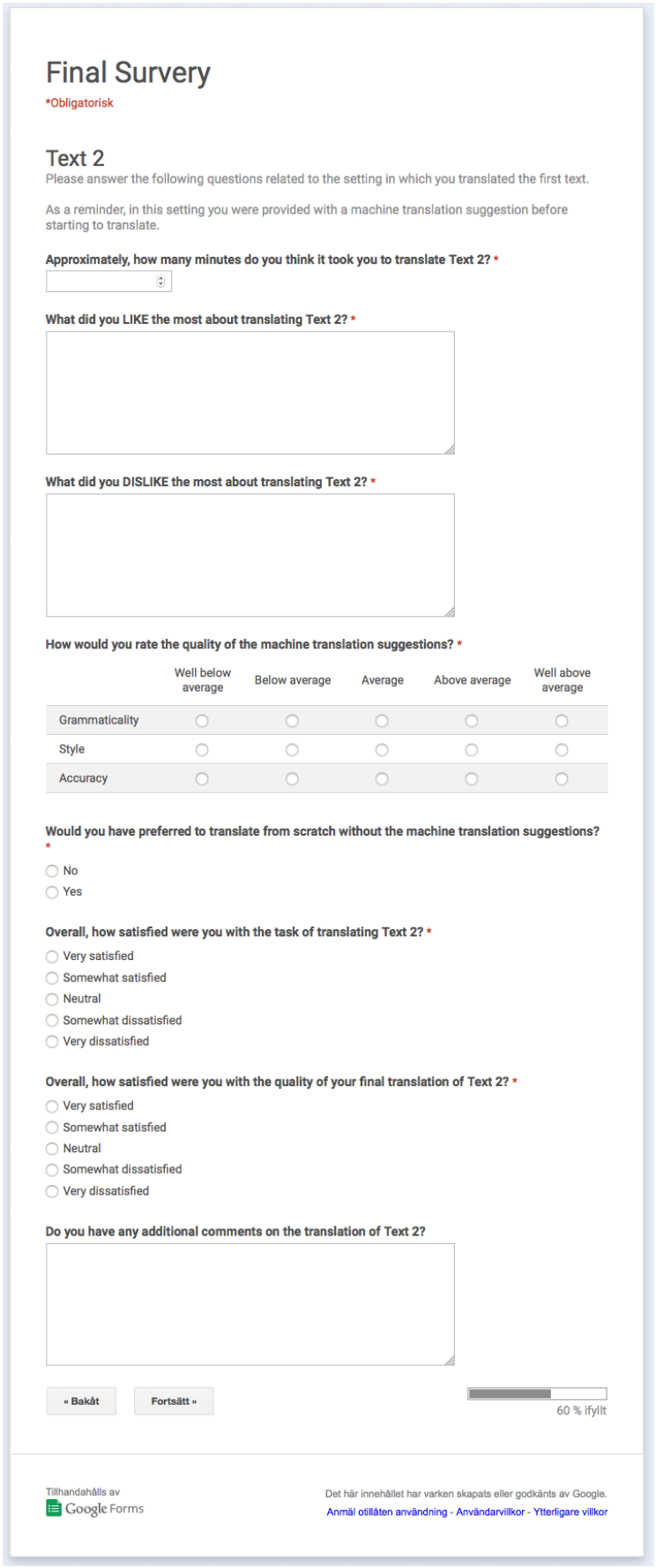


Figure 29: Final Questionnaire (Page 2)



Final Survery

*Obligatorisk

Text 2

Please answer the following questions related to the setting in which you translated the first text.

As a reminder, in this setting you were provided with a machine translation suggestion before starting to translate.

Approximately, how many minutes do you think it took you to translate Text 2? *

What did you LIKE the most about translating Text 2? *

What did you DISLIKE the most about translating Text 2? *

How would you rate the quality of the machine translation suggestions? *

	Well below average	Below average	Average	Above average	Well above average
Grammaticality	○	○	○	○	○
Style	○	○	○	○	○
Accuracy	○	○	○	○	○

Would you have preferred to translate from scratch without the machine translation suggestions? *

- ○ No
- ○ Yes

Overall, how satisfied were you with the task of translating Text 2? *

- ○ Very satisfied
- ○ Somewhat satisfied
- ○ Neutral
- ○ Somewhat dissatisfied
- ○ Very dissatisfied

Overall, how satisfied were you with the quality of your final translation of Text 2? *

- ○ Very satisfied
- ○ Somewhat satisfied
- ○ Neutral
- ○ Somewhat dissatisfied
- ○ Very dissatisfied

Do you have any additional comments on the translation of Text 2?

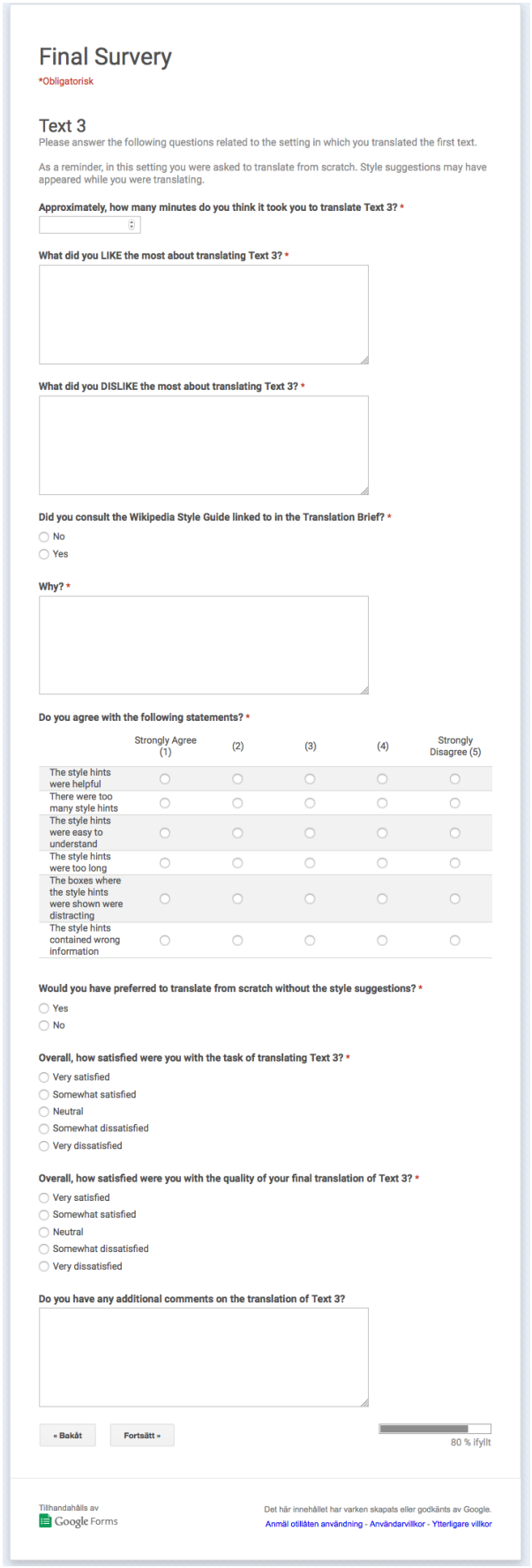
« Bakåt Fortsätt »

60 % ifyllt

Tillhandahålls av Google Forms

Det här innehållet har varken skapats eller godkänts av Google.
Anmäl otillåten användning - Användarvillkor - Ytterligare villkor

Figure 30: Final Questionnaire (Page 3)



Final Survery

*Obligatorisk

Text 3

Please answer the following questions related to the setting in which you translated the first text.

As a reminder, in this setting you were asked to translate from scratch. Style suggestions may have appeared while you were translating.

Approximately, how many minutes do you think it took you to translate Text 3? *

What did you LIKE the most about translating Text 3? *

What did you DISLIKE the most about translating Text 3? *

Did you consult the Wikipedia Style Guide linked to in the Translation Brief? *

- No
- Yes

Why? *

Do you agree with the following statements? *

	Strongly Agree (1)	(2)	(3)	(4)	Strongly Disagree (5)
The style hints were helpful					
There were too many style hints					
The style hints were easy to understand					
The style hints were too long					
The boxes where the style hints were shown were distracting					
The style hints contained wrong information					

Would you have preferred to translate from scratch without the style suggestions? *

- Yes
- No

Overall, how satisfied were you with the task of translating Text 3? *

- Very satisfied
- Somewhat satisfied
- Neutral
- Somewhat dissatisfied
- Very dissatisfied

Overall, how satisfied were you with the quality of your final translation of Text 3? *

- Very satisfied
- Somewhat satisfied
- Neutral
- Somewhat dissatisfied
- Very dissatisfied

Do you have any additional comments on the translation of Text 3?

« Bakåt Fortsätt »

80 % ifyllt

Tillhandahålls av Google Forms

Det här innehållet har varken skapats eller godkänts av Google.
Anmäl otillåten användning - Användarvillkor - Ytterligare villkor

Figure 31: Final Questionnaire (Page 4)

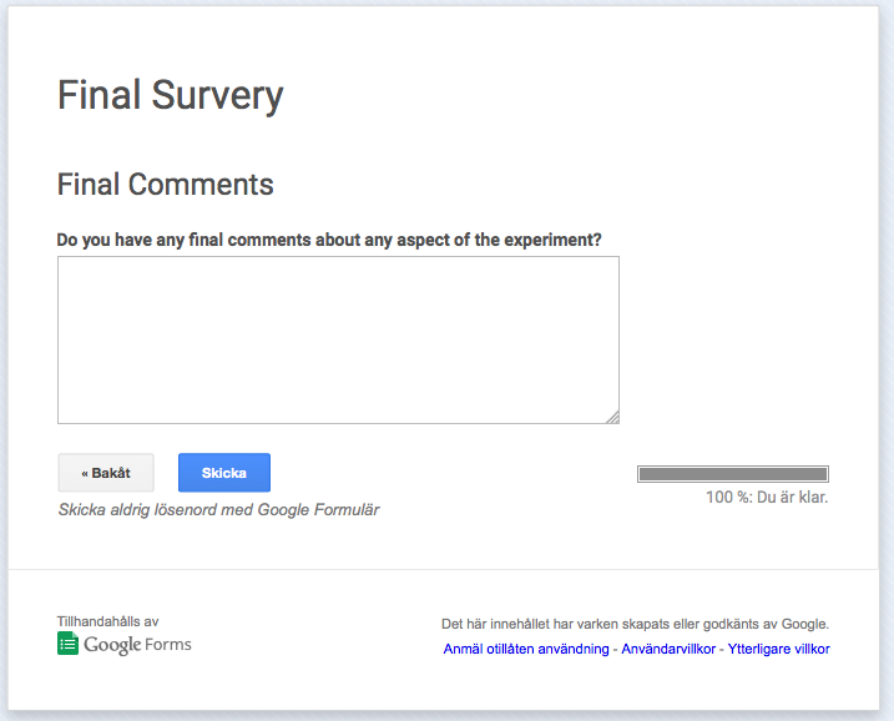


Figure 32: Final Questionnaire (Page 5)

C

WIKIPEDIA TEXTS

C.1 SOURCE TEXTS

The following subsections contain the Wikipedia texts that participants were asked to translate. All links to other Wikipedia articles were removed and not shown to participants, as it was considered too much work for this experiment to ensure that the translation contained the correct links in Spanish.

C.1.1 *History of Artificial Intelligence*

> The field of AI research was founded at a conference on the campus of Dartmouth College in the summer of 1956.
> [...]
> In 1973 [...] the U.S. and British Governments stopped funding undirected research into artificial intelligence. Seven years later, a visionary initiative by the Japanese Government inspired governments and industry to provide AI with billions of dollars, but by the late 80s the investors became disillusioned and withdrew funding again. This cycle of boom and bust, of "AI winters" and summers, continues to haunt the field.
>
> (Wikipedia, 2015c)

C.1.2 *Charlotte Gyllenhammer*

A slight motification was introduced to this text in order to include a style error. The Wikipedia Spanish Style Guide (Wikipedia, 2015d) states that lists which can be ordered chronologically, should be ordered in this fashion. Although the original text correctly ordered the artist's works ("*Belle*, 1998, *Disobedience*, 1998, *Fall*, 1999, and more recently *Hang* 2006"), the order of *Disobedience* and *Fall* was swapped, making a clearly unordered list. Also, the bold font styling of "**Charlotte Gyllenhammer**" was removed as it was deemed there were sufficient other style guide considerations in the text.

> Charlotte Gyllenhammar, born 1963, is a fine artist based in Stockholm, Sweden. She began her career as a painter, but swiftly moved on to sculpture and installation after completing her studies at the Royal College of Art in London.

> [...]
> The work entitled *Die for You* was the first step in a progression of images and environments that invert perspective. For example, confinement and inversion are evident in her video/photographic series of suspended women entitled *Belle*, 1998, *Fall*, 1999,*Disobedience*, 1998, and more recently *Hang* 2006.
>
> (Wikipedia, 2015a)

C.1.3 *Garfield*

The first sentence of the text was shortened from the original, since the long length of the paragraph slowed down the StyleCheck tool and could even block it. The original read "[...] any form of exertion; his constant shedding (which constantly annoys Jon); and his abuse of Odie and Jon as well as his obsession with mailing Nermal to Abu Dhabi."

> Many of the gags focus on Garfield's obsessive eating and obesity; his fear of spiders (many of these can be found in the 7th strip comic collection type book); his hate of Mondays, diets, and any form of exertion; and his obsession with mailing Nermal to Abu Dhabi. Though he will eat nearly anything (with the exception of raisins and spinach), Garfield is particularly fond of lasagna; he also enjoys eating Jon's houseplants and other pets (mainly birds and fish).
>
> (Wikipedia, 2015b)

C.2 MT TRANSLATION SUGGESTIONS

Machine translation suggestions were obtained by running the text fragments through Google Translate[1]. Below are the results as they were used in the PE setup of the experiments.

C.2.1 *History of Artificial Intelligence (MT)*

> El campo de la investigación en IA fue fundada en una conferencia en el campus de la universidad de Dartmouth en el verano de 1956.
>
> En 1973 los EE.UU. y los gobiernos británicos detuvimos financiación de la investigación no dirigida en inteligencia artificial. Siete años más tarde, una iniciativa

1 https://translate.google.com/, translations were carried out on 4 May 2015.

visionaria de los gobiernos Gobierno inspirado japonés y la industria para proporcionar AI con miles de millones de dólares, pero a finales de los años 80 los inversores se desilusionó y se retiró la financiación de nuevo. Este ciclo de auge y caída, de los "inviernos y veranos AI", sigue acosando el campo.

C.2.2 *Charlotte Gyllenhammer (MT)*

Charlotte Gyllenhammar, nacido en 1963, es un artista plástico con sede en Estocolmo, Suecia. Comenzó su carrera como pintor, pero rápidamente se trasladó a la escultura y la instalación después de completar sus estudios en el Royal College of Art de Londres.

El trabajo titulado Die for You fue el primer paso en una progresión de imágenes y ambientes que invierten perspectiva. Por ejemplo, el parto y la inversión son evidentes en su video / serie fotográfica de las mujeres en suspensión con derecho Belle, 1998, Otoño, 1999, Desobediencia, 1998, y más recientemente Cuelgue 2006.

C.2.3 *Garfield (MT)*

Muchos de los gags se centran en la alimentación y la obesidad obsesivo de Garfield; su miedo a las arañas (muchos de ellos se puede encontrar en el séptimo tira tipo de colección cómica del libro); su odio de los lunes, dietas, y cualquier forma de ejercicio; y su obsesión por correo Nermal a Abu Dhabi. A pesar de que va a comer casi cualquier cosa (con la excepción de las pasas y las espinacas), Garfield es particularmente aficionado a la lasaña; también le gusta comer las plantas de interior de Jon y otros animales (principalmente aves y peces).

D

STYLECHECK GRAMMAR

D.1 WIKI ABSTRACT SYNTAX

Listing 2: Abstract syntax where the style rules are defined

```
1  --# -path=.:alltenses:../chunk:../style:../translator
2
3  abstract Wiki = Cat, Chunk, StyleCatAbs
4  -- Avoid generating parses for both SymbS and S_Chunk, which
       complicates the grammar.
5  , Symbol [Symb, MkSymb]
6  -- For the rules and hints:
7  , Season, List, PunctuationAbs, PresentReference
8  ** {
9
10 flags startcat = Phr ;
11
12 -- CHUNKS
13 fun RuleChunk : StyleRule -> Chunk ;
14 fun HintChunk : StyleHint -> Chunk ;
15 fun LookupChunk : StyleLookup -> Chunk ;
16
17 -- TEXT 1: History of Artificial Intelligence
18 ---- 1.A. Seasons
19 fun AvoidSeasons_Hint : Season -> StyleHint ;
20 ---- 1.B. Quotes
21 fun Quote_Rule : Phr -> StyleRule ;
22 ---- 1.C. Numbers in words or figures (Not rendered in the Style
       Grammar)
23 ---- 1.D. Capitalise
24 fun Capitalise_Hint : StyleHint ;
25 ---- 1.E. Acronyms
26 fun USA_Rule : StyleRule ;
27 ---- 1.F. Historic Present (Not rendered in the Style Grammar)
28
29
30 -- TEXT 2: Charlotte Gyllenhammar
31 ---- 2.A. Lists in chronological order
32 fun Chronological_Lists_Hint : PhrList -> StyleHint ;
33 ---- 2.B. Avoid deictic and anaphoric expressions that refer to
       the time of writing
34 fun Present_Reference_Hint : PresentRef -> StyleHint ;
35 ---- 2.C. Slash sign optionality
36 fun Slash_Optionality_Lookup : StyleLookup ;
37
38 -- TEXT 3: Garfield
```

```
39 ---- 3.A. Excessive digressions
40 fun Digressions_Hint : Phr -> StyleHint ;
41 ---- 3.B. Ordinal abbreviation
42 fun Ordinal_Abbreviation_Rule : StyleRule ;
43 ---- 3.C. Numbers in words or figures (Not rendered in the Style
       Grammar)
44 ---- 3.D. Abu Dabi
45 fun Abu_Dabi_Rule : StyleRule ;
46 }
```

D.2 WIKIMASTER CONCRETE SYNTAX

Listing 3: Main concrete syntax where the style records are built

```
1  --# -path=.:alltenses:../chunk:../style:../translator
2
3  concrete WikiMaster of Wiki =
4  SymbolSpa [MkSymb, Symb], ChunkSpa, StyleCat
5  -- For the rules and hints:
6  , SeasonSpa, ListSpa, PresentReferenceSpa
7  ** open Prelude, ParadigmsSpa, SyntaxSpa, StyleRules
8  -- For the rules and hints:
9  , ManualDeEstilo, Typography, EntitiesSpa
10 in {
11
12 -- CHUNKS
13 lin RuleChunk r = r.options ;
14 lin HintChunk h = h.options ;
15 lin LookupChunk l = l.options ;
16
17 -- TEXT 1: History of Artificial Intelligence
18 ---- 1.A. Seasons
19 lin AvoidSeasons_Hint s = mkStyleHint (mkNP s) W_11_Estaciones ;
20 ---- 1.B. Quotes
21 lin Quote_Rule p = mkStyleRule guillemet_Surround (double_
       straight_Surround | double_curly_Surround | single_straight_
       Surround) p W_12_Comillas ;
22 ---- 1.C. Numbers in words or figures (Not rendered in the Style
       Grammar)
23 ---- 1.D. Capitalise
24 lin Capitalise_Hint = mkStyleHint (mkN ((optionalInitCap "
       gobierno") | (optionalInitCap "gobiernos"))) DPD_May sculas
       _4_28 ;
25 ---- 1.E. Acronyms
26 lin USA_Rule = mkStyleRule ("EUA" | "EE. UU.") ("EE.UU." | "USA"
       | "U.S.A.") DPD_EUA ;
27 ---- 1.F. Historic Present (Not rendered in the Style Grammar)
28
29 -- TEXT 2: Charlotte Gyllenhammar
30 ---- 2.A. Lists in chronological order
```

```
31 lin Chronological_Lists_Hint l = mkStyleHint l W_23_Chronological
       _lists ;
32 ---- 2.B. Avoid deictic and anaphoric expressions that refer to
       the time of writing
33 lin Present_Reference_Hint r = mkStyleHint r W_17_Deictic_
       expressions ;
34 ---- 2.C. Slash sign optionality
35 lin Slash_Optionality_Lookup = mkStyleLookup (mkStyleHint forward
       _slash_Str DPD_Barra_b) ;
36
37 -- TEXT 3: Garfield
38 ---- 3.A. Excessive digressions
39 lin Digressions_Hint p = mkStyleHint (brackets_Surround | square_
       brackets_Surround | em_dash_Surround | en_dash_Surround) p W
       _2_Digressions ;
40 ---- 3.B. Ordinal abbreviation
41 lin Ordinal_Abbreviation_Rule = mkStyleRule ("7.  " | "s ptimo")
        (spaceOrdered "7" "  " | spaceOrdered "7" "o" | dotOrdered
       "7" "o") W_Abr_Puntos_Volados ;
42 ---- 3.C. Numbers in words or figures (Not rendered in the Style
       Grammar)
43 ---- 3.D. Abu Dabi
44 lin Abu_Dabi_Rule = mkStyleRule (mkN "Abu Dabi") ((mkN "Abu Dahbi
       ") | (mkN "Abu Dhabi")) DPD_Abu_Dabi ;
45
46 oper
47 -- Opers to help in creating various types of alternatives
48     -- OneTwo or OneMiddleTwo
49     glueOrJoin : Str -> Str -> Str -> Str = \s1,s2,middle -> (s1
           + s2 | s1 + middle + s2) ;
50     -- OneTwo or One Two
51     spaceOrdered : Str -> Str -> Str = \s1,s2 -> glueOrJoin s1 s2
            " " ;
52     -- OneTwo, One Two, TwoOne or Two One
53     spaceUnordered : Str -> Str -> Str = \s1,s2 -> ((spaceOrdered
            s1 s2) | (spaceOrdered s2 s1)) ;
54     -- OneTwo or One.Two
55     dotOrdered : Str -> Str -> Str = \s1,s2 -> glueOrJoin s1 s2
           "." ;
56     -- one or One
57     optionalInitCap : Str -> Str = \s -> (s | (toUpperFirst s)) ;
58 } ;
```

D.3 WIKIHINT CONCRETE SYNTAX

Listing 4: Concrete syntax where the hints are selected from the style records

```
1 --# -path=.:alltenses:../chunk:../style:../translator
2 concrete WikiHint of Wiki = WikiMaster - [RuleChunk, HintChunk,
      LookupChunk] ** {
```

```
3
4 lin RuleChunk r = r.hint ;
5 lin HintChunk h = h.hint ;
6 lin LookupChunk l = l.hint ;
7 }
```

BIBLIOGRAPHY

Alves, Fabio (2003). *Triangulating Translation: Perspectives in Process Oriented Research*. John Benjamins. Amsterdam, Philadelphia (Cited on page 10).

Angelov, Krasimir and Aarne Ranta (2009). "Implementing controlled languages in GF." In: *CNL'09: Proceedings of the 2009 conference on Controlled natural language*. Springer-Verlag, pp. 82–101 (Cited on page 28).

Brooke, John (1996). "SUS-A quick and dirty usability scale." In: *Usability evaluation in industry* 189.194, pp. 4–7 (Cited on page 41).

Carl, Michael (2012). "The CRITT TPR-DB 1.0: A Database for Empirical Human Translation Process Research." In: *AMTA 2012 Workshop on Post-Editing Technology ...* (Cited on pages 11, 61).

Carl, Michael, Silke Gutermuth, and Silvia Hansen-Schirra (2015). "Post-Editing Machine Translation: Efficiency, Strategies, and Revision Processes in Professional Translation Settings." In: *Psycholinguistic and Cognitive Inquiries into Translation and Interpreting*. Ed. by Aline Ferreira and John W. Schwieter. 2015; 7. John Benjamins Publishing Company, pp. 145–174. ISBN: 9789027258557 (Cited on pages 15, 19, 24, 41, 46, 54, 56).

Catford, John Cunnison (1965). *A linguistic theory of translation*. Vol. 74. Oxford University Press London (Cited on page 5).

Christensen, Tina Paulsen (2011). "Studies on the mental processes in translation memory-assisted translation–The state of the art." In: *trans-kom. Zeitschrift für Translationswissenschaft und Fachkommunikation* 4.2, pp. 137–160 (Cited on pages 10, 14, 40).

Christensen, Tina Paulsen and Anne Schjoldager (2011). "The Impact of Translation-Memory (TM) Technology." In: *Human-Machine Interaction in Translation: Proceedings of the 8th International NLPCS Workshop*. Vol. 41. Samfundslitteratur, p. 119 (Cited on pages 14, 19).

Chuang, Thomas C et al. (2005). "Collocational translation memory extraction based on statistical and linguistic information." In: *International Journal of Computational Linguistics and Chinese Language Processing* 10.3, pp. 329–346 (Cited on page 14).

Čulo, Oliver (2014). "Approaching Machine Translation from Translation Studies -– a perspective on commonalities, potentials, differences." In: *Proceedings of the 17th Annual Conference of EAMT* (Cited on page 9).

Dragsted, Barbara (2004). *Segmentation in translation and translation memory systems: An empirical investigation of cognitive segmentation*

and effects of integrating a TM system into the translation process (Cited on page 14).

Dragsted, Barbara (2006). "Computer-aided translation as a distributed cognitive task." In: *Pragmatics & Cognition* 14.2, pp. 443–464 (Cited on page 14).

Española, Real Academia and Asociación de Academias de la Lengua Española (2005). *Diccionario panhispánico de dudas*. Real Academia Española (Cited on page 29).

Esselink, Bert (2000). "From translation to localisation — and back." In: *Translating Science: 2nd International Conference on Specialized Translation*. Ed. by J. Chabás et al. Universitat Pompeu Fabra (Cited on page 8).

Federico, Marcello, Alessandro Cattelan, and Marco Trombetti (2012). "Measuring user productivity in machine translation enhanced computer assisted translation." In: *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)* (Cited on pages 48, 49).

Federico, Marcello et al. (2014). "The MateCat tool." In: *Proceedings of COLING*, pp. 129–132 (Cited on page 14).

Hardmeier, Christian (2014). "Discourse in Statistical Machine Translation." PhD thesis. Uppsala University, Department of Linguistics and Philology, p. 185 (Cited on page 19).

Holz-Mänttäri, Justa and Suomalainen Tiedeakatemia (1984). *Translatorisches Handeln: Theorie und Methode*. Suomalainen tiedeakatemia Helsinki (Cited on page 6).

Hutchins, Edwin (2000). "Distributed cognition." In: *International Encyclopedia of the Social and ...* (Cited on page 10).

Jiménez-Crespo, Miguel A (2010). "The effect of Translation Memory tools in translated Web texts: Evidence from a comparative product-based study." In: *Linguistica Antverpiensia* 8, pp. 213–232 (Cited on page 14).

Kaljurand, Kaarel and Tobias Kuhn (2013). "A Multilingual Semantic Wiki Based on Attempto Controlled English and Grammatical Framework." In: *The Semantic Web: Semantics and Big Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 427–441 (Cited on page 28).

Kay, Martin (1980). "The Proper Place of Men and Machines in Language Translation." In: (Cited on page 3).

— (1997). "It's Still the Proper Place." English. In: *Machine Translation* 12.1/2, pp. 35–38 (Cited on pages 3, 8).

Koehn, Philipp et al. (2007). "Moses: Open source toolkit for statistical machine translation." In: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, pp. 177–180 (Cited on page 3).

Koponen, Maarit et al. (2012). "Post-editing time as a measure of cognitive effort." In: *Proceedings of WPTP* (Cited on page 49).

Krings, Hans Peter (2005). "Wege ins Labyrinth–Fragestellungen und Methoden der Übersetzungsprozessforschung im Überblick." In: *Meta: Journal des traducteursMeta:/Translators' Journal* 50.2, pp. 342–358 (Cited on page 40).

Lagoudaki, Elina (2006). "Translation memories survey 2006: Users' perceptions around tm use." In: *proceedings of the ASLIB International Conference Translating & the Computer*. Vol. 28. 1, pp. 1–29 (Cited on page 13).

Lewis, Philip E. (1985). "The Measure of Translation Effects." In: *Difference in Translation*. Ed. by J. F. Graham. Ithaca NY: Cornell University Press (Cited on page 7).

Nord, Christiane (1997). "Translation theories explained: translating as a purposeful activity." In: *Manchester: St. Jerome* (Cited on page 6).

O'Brien, Sharon (2002). "Teaching post-editing: a proposal for course content." In: *6th EAMT Workshop Teaching Machine Translation* (Cited on page 15).

— (2009). "Eye tracking in translation process research: methodological challenges and solutions." In: *Methodology, Technology and Innovation in Translation Process Research* 38, pp. 251–266 (Cited on page 11).

— (2012). "Translation as human–computer interaction." In: *Translation Spaces* 1.1, pp. 101–122 (Cited on page 19).

Papineni, Kishore et al. (2001). "BLEU." In: *the 40th Annual Meeting*. Morristown, NJ, USA: Association for Computational Linguistics, pp. 311–318 (Cited on page 50).

Pierce, John R. and John B. Carroll (1966). *Languages and Machines — Computers in Translation and Linguistics*. Tech. rep. Washington, DC: Automatic Language Processing Advisory Committe (ALPAC), National Academy of Sciences (Cited on page 3).

Pym, Anthony (2009). *Exploring translation theories*. Routledge (Cited on pages 5, 7–9).

Quine, Willard Van Orman (1969). "Linguistics and Philosophy." In: *Language and Philosophy: A Symposium*. Ed. by Sydney Hook. New York: New York University Press (Cited on page 7).

Ranta, Aarne (2009). "The GF Resource Grammar Library." In: *Linguistics in Language Technology* 2.2, pp. 1–65 (Cited on page 28).

— (2011). *Grammatical framework: Programming with multilingual grammars*. CSLI Publications, Center for the Study of Language and Information (Cited on page 28).

— (2014). "Embedded Controlled Languages." In: *arXiv.org*. arXiv: 1406.4057v1 [cs.CL] (Cited on pages 28, 31).

Reiß, Katharina and Hans J Vermeer (1984). *Grundlegung einer allgemeinen Translationstheorie*. Vol. 147. Walter de Gruyter (Cited on page 6).

Risku, Hanna (2010). "A cognitive scientific view on technical communication and translation: Do embodiment and situatedness really make a difference?" In: *Target* 22.1, pp. 94–111 (Cited on page 10).

Snell-Hornby, Mary (1988). *Translation studies: an integrated approach*. John Benjamins Publishing (Cited on page 7).

Toury, Gideon (1995a). "The Nature and Role of Norms in Translation." In: *Descriptive Translation Studies and Beyond*. Amsterdam, Philadelphia: John Benjamins (Cited on page 6).

— (1995b). "The Notion of 'Assumed Translation' – An Invitation to a New Discussion." In: *Letterlijkheid, Woordelijheid / Literality, Verbality*. Ed. by Henri Bloemen, Erik Hertog, and Winibert Segers. Antwerpen and Hermelen: Fantom (Cited on page 6).

Vidal Hussey, Daniel (2013). "La guia d'estil aplicada digitalment." Unpublished Bachelor's Thesis (Cited on pages 27, 30).

Washbourne, Kelly (2012). "Translation style guides in translator training: Considerations for task design." In: *The Journal of Specialised Translation* 17 (Cited on page 27).

Wikipedia (2015a). *Charlotte Gyllenhammar — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-April-2015]. URL: http://en.wikipedia.org/w/index.php?title=Charlotte_Gyllenhammar&oldid=647293586 (Cited on page 90).

— (2015b). *Garfield — Wikipedia, The Free Encyclopedia*. [Online; accessed 15-April-2015]. URL: http://en.wikipedia.org/w/index.php?title=History_of_artificial_intelligence&oldid=663822581 (Cited on page 90).

— (2015c). *History of artificial intelligence— Wikipedia, The Free Encyclopedia*. [Online; accessed 15-April-2015]. URL: http://en.wikipedia.org/w/index.php?title=History_of_artificial_intelligence&oldid=663822581 (Cited on page 89).

— (2015d). *Wikipedia: Manual de Estilo — Wikipedia, The Free Encyclopedia*. [Online; accessed 2-April-2015]. URL: http://es.wikipedia.org/wiki/Wikipedia:Manual_de_estilo (Cited on page 89).

Zabalbeascoa, Patrick (1999). "Priorities and restrictions in translation." In: *Translation and the (Re) Location of Meaning. Leuven: CETRA*, pp. 159–167 (Cited on pages 57, 62).

LIST OF FIGURES

LIST OF TABLES

LISTINGS

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*.

Final Version as of September 14, 2015 (`classicthesis` Version 1.0).