Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Traduzione, interpretazione e interculturalità

Ciclo XXXII

Settore Concorsuale: 10/L1

Settore Scientifico Disciplinare: L-LIN/12

MACHINE TRANSLATION FOR INSTITUTIONAL ACADEMIC TEXTS: OUTPUT QUALITY, TERMINOLOGY TRANSLATION AND POST-EDITOR TRUST

Presentata da: Randy Scansani

Coordinatore Dottorato

Raffaella Baccolini

Supervisore

Silvia Bernardini

Supervisore

Luisa Bentivogli

Fondazione Bruno Kessler

Esame finale anno 2020

Contents

Chapter 1

Introduction and research questions

1.1 Introduction

Following the Bologna process, universities have been urged to increase their degree of internationalisation, with the aim of creating a European Higher Education Area (EHEA) that encourages students' mobility. This process has brought with it the need of communicating effectively in English also for institutions based in countries where this is not an official language. As one of the aims in the creation of the EHEA was to foster students' mobility, availability of multilingual course unit descriptions (or course catalogues) has become especially important.

However, previous work has shown that institutional academic communication has not undergone a substantial increase of translated content, both from a qualitative and from a quantitative point of view. Callahan and Herring (2012) claim that the number of universities whose website contents are translated into English varies across the European Union, with Northern and Western countries paying more attention to their internationalisation than Southern ones. When quality is in focus, things do not improve: many of the translated documents feature terminological inconsistencies (Candel-Mora and Carrió-Pastor, 2014). This is due to the absence of standardised terminological resources that could assist in the drafting of institutional academic texts. Terminology is indeed a key factor for the production of such texts, since they usually feature terms that are typical of institutional academic communication, but also expressions that belong to the discipline taught (Ferraresi, 2017). Such peculiarities make these texts an interesting case study. The issues they pose make course catalogues an ideal test bed for a number of tasks related to machine translation (MT). On the other hand, adopting MT would offer universities the opportunity to streamline their translation process.

In the present work, institutional academic texts are exploited to contribute to research in the field of MT, focusing on two main research questions.

1. ***Can MT be profitably applied to the translation of institutional academic texts?*** The use of tools, and especially of machine translation (MT) systems, supporting translation in the institutional academic domain would be beneficial for universities, helping to handle the large number of course unit descriptions and degree programme descriptions that have to be produced on a yearly basis through a mix of drafting from scratch and partial revisions or updates. Exploring the feasibility of successfully applying MT to the translation of institutional academic texts poses one particular challenge. MT systems usually struggle when confronted with texts featuring highly-specialised terminology, as testified by the amount of work invested to inject terminology into different MT archi-

tectures (Arcan et al., 2014b; Bouamor et al., 2012; Chatterjee et al., 2017), and neural machine translation (NMT), which currently is the state-of-the-art MT architecture (Bahdanau et al., 2014; Bentivogli et al., 2016; Vaswani et al., 2017a), is no exception to this rule (Hasler et al., 2018). Institutional academic texts feature multi-domain terminology. Understanding how the ability of correctly translating terms can be enhanced and evaluated is thus of the essence here and can provide interesting input for future research on MT and terminology in other domains.

The present work contributes to this field of research in the following ways. First, German–English and Italian–English data sets are created, and MT engines are trained and tested in the institutional academic domain. The choice of Italian and German offers insights into a Romance language and a Germanic one. Then, terms in the test set are manually annotated creating a gold standard used to assess terminology translation. Given the time required by the development of a new pipeline to annotate terms and build the data set, this part of the work focuses on Italian–English only. To conclude, the feasibility of applying MT to institutional academic texts is investigated by evaluating the output quality obtained for both language combinations with automatic metrics, and by evaluating term translation for Italian–English exploiting the manually annotated terminology gold standard.

2. ***Do translators trust MT?*** The feasibility of profitably applying MT to institutional academic texts does not only depend on the output quality, but also on the willingness of the final users to work with the provided MT output. Possible preconceptions, e.g. translators lack of trust towards the text they are asked to post-edit might hinder the benefits of the application of MT. While being investigated in several fields (Blomqvist, 1997; McKnight and Chervany, 2001), to the best of my knowledge user trust towards the MT output has been neglected so far, except for the work by Martindale and Carpuat (2018). The negative opinion of current professional translators towards MT has been pointed out (Läubli and Orrego-Carmona, 2017) together with the reasons for their choice not to adopt MT suggestions (Cadwell et al., 2018). However, in the recent past several degree programmes in translation have started to offer courses on MT and post-editing, while advances in MT architecture have brought improvements in the output quality. It is thus likely that the next generation of translators will enter the market with a different opinion on MT and the post-editing activity.

To answer this research question, an experiment to measure participants trust towards an MT output was structured, providing insights into this psychological aspect influencing MT adoption.

1.2 Method

1.2.1 Can MT be profitably applied to the translation of institutional academic texts?

The implementation of MT systems in universities' translation pipeline is arguably a necessary step in the process of streamlining multilingual communication in the institutional academic world. As previously introduced (see Sect. 1.1), bilingual course unit descriptions and degree programme descriptions are collected for Italian–English and German–English, with which an in-domain corpus is built for each language combination. After the data collection step, corpora are split into one data set that is used to train the MT

engine and one test set used to assess the quality achieved.

The two engines used in each scenario are the free generic Google Translate (GT)[1] and the commercial version of ModernMT (MMT).[2] Both systems were chosen for being state-of-the-art ready to use NMT systems. Moreover, MMT implements an adaptation mechanism.

Evaluations, using automatic metrics, are carried out in order to assess the engines' ability to handle institutional academic texts despite their different levels of terminology standardisation and the small amount of available sentence pairs. Tests are carried out in two scenarios, the first being an entry-level one where no bilingual sentences are available, the second one where a reasonable amount of bilingual sentences is available to perform domain adaptation. For Italian–English, multi-domain terminology was manually annotated in a test set of ca. 2.000 sentence pairs to also allow for an in-depth assessment of terminology translation of the engines in the different scenarios. The annotated data set, called MAGMATic (Multi-domain Academic Gold Standard with Manual Annotation of Terminology) is released under a Creative Commons Attribution – Non Commercial – Share Alike 4.0 International license (CC BY-–NC–SA 4.0) to contribute to future research on terminology translation.[3]

1.2.2 Do translators trust MT?

In a second step, we investigate possible preconceptions of translator trainees towards the MT output and how these might affect productivity. As stated above (see Sect. 1.1), assessing the willingness of post-editors to work on an MT output is an essential part of the process that verifies the applicability of MT to a particular text domain.

Simulating a real-world professional task, students from the Master's Degree in Specialised Translation of the University of Bologna are assigned the same MT output of two course unit descriptions.[4] Half of them are asked to post-edit it, while the other half is told that the text was translated by a human and is in need of revision. Participants work in the free online CAT tool MateCat[5] where a project including the target text and a termbase is assigned to each of them.

Participants' productivity in the task is compared based on the number of edits performed – using HTER (Snover et al., 2006) – and based on temporal effort – words per second (WPS). In the two tasks, we aim at finding a possible bias caused by different levels of trust towards the two translation methods.

1.3 Research questions: summary

The present thesis is structured as follows. First, a background on the application scenario, i.e. on institutional academic texts, is provided in Chapter 2. A review of the literature is provided for the main topics explored, i.e. MT (Chapter 3), and the concept of trust (Chapter 4) – this last one being particularly relevant since trust has been rarely confronted in MT research.

[1] `https://translate.google.com/`
[2] `https://www.modernmt.eu/`
[3] `https://ict.fbk.eu/magmatic/`
[4] For more information on the degree programme: `https://bit.ly/2nXKN8S`
[5] `www.matecat.com`

Following this overview, the main parts of the present thesis are introduced in two distinct chapters, each answering to one of the two research questions mentioned above (see Sect. 1.1). The chapter on the application of MT to institutional academic texts (Chapter 5) starts from the method followed to collect bilingual texts and build parallel corpora for German–English and Italian–English. Then, the output quality obtained leveraging this corpora is analysed in different scenarios and different resource settings. The assessment finally concentrates on Italian–English with a focus on terminology translation. To introduce this last part, the manual annotation process used to build MAGMATic is detailed.

The chapter on translators' trust towards MT (Chapter 6) has three main parts. In the first part, motivations for this experiment are outlined, then the main goals and variables are introduced. Following these sections, structure and results of a pilot experiment taking place in May 2018 are summarised. Given the issues spotted in this pilot, the structure was modified leading to a second experiment (March 2019) whose structure and results are analysed.

The concluding part of the present thesis (Chapter 7) discusses the main achievements together with their limitations and their input for future research in the field. Given the aim of training an MT engine to streamline the translation of institutional academic texts, this discussion focuses not only on the contribution for the research world, but also on the practical contribution to institutional academic communication.

Chapter 2

Institutional academic communication

2.1 Institutional academic texts and the Bologna Process

The end of the 20th century has brought societal, institutional and economic changes to Europe and all over the world. Globalisation, integration of economies and the wave of new technologies have been supported by – and at the same time have encouraged – actions and initiatives by institutions all over the world to promote common policies in several fields. These shared practices have boosted the free movement of people, and for companies and institutions the capability of attracting people from other countries has become an added value. Clearly, the increasing interest in internationalisation has had a major impact on the way in which foreign languages and translation were conceived, taught and exploited, with a significant growth in the demand for translations. It has to be noticed that in this work the concept of internationalisation is referred to as the ability to effectively communicate in English in order to attract investments, stakeholders or workers from abroad.

In the European higher education field, the inception of the Bologna process has paved the way for the creation of a European Higher Education Area (EHEA) that encourages integration of education systems, with the ultimate aim of creating a single one able to foster students' and staff's mobility. EHEA declared as its goals "increasing the mobility of students and staff thanks to common concrete tools such as European credits transfer system (ECTS), structuration of the studies into three cycles and quality assurance of higher education".[1] This process has brought with it the need of communicating effectively in English also for institutions based in countries where this is not an official language. This is particularly true for academic websites, which are a source of information for 84% of prospective students and are also the most prominent source of institutional academic texts in general – including course unit descriptions, mission statements, announcements, etc. (Ferraresi and Bernardini, 2013).

Despite the ambitious vision set up by the EHEA members – and despite the growing use of English as a *lingua franca* in the higher education domain –, Callahan and Herring (2012) have shown that institutional academic communication has neither undergone a substantial increase of translated content from a quantitative point of view, nor has it shown sizeable improvements from a qualitative point of view. The authors claim that the number of universities whose website contents are translated into English varies across the European Union (EU), with North-Western countries paying more attention to

[1] Official website of the 2018 EHEA Ministerial Conference, consulted in August 2018. `http://www.ehea2018.paris/`

their internationalisation than Southern ones. When the focus is on the quality of available English contents, things do not improve: many of the translated documents feature terminological inconsistencies (Candel-Mora and Carrió-Pastor, 2014). Moreover, the lack of terminology harmonisation is a well known issue for institutional texts, which has also been acknowledged by EU institutions (Crosier, Purser, and Smidt, 2007, p. 20) and has brought to the creation of European glossaries and termbases such as the Education section of the IATE database[2] or the Eurydice glossaries.[3] However, the effort of many universities to develop self-built terminology resources proves that the approach adopted by EU institutions has not been successful (Ferraresi, 2017).

2.2 The case of course catalogues

Among multilingual web contents, course catalogues are particularly key for prospective students, since they contain detailed descriptions of the course contents. According to the ECTS guide (p. 55)[4] a course catalogue must contain – among others – texts describing the institution, academic authorities, academic calendar, list of programmes offered, admission requirements, ECTS credit allocation policy, information on programmes and information on individual educational components. Two of these descriptions are relevant for the present work. Course unit descriptions (CUDs) provide information on individual educational components (see Sect. 2.2.2).[5] Degree programme descriptions (DPDs) outline information on degree programmes (see Sect. 2.2.1).[6]

The translation of such texts is particularly relevant not only for (prospective) students, but also for universities in many ways. First, according to the ECTS users' guide, their translation into English is required for higher education institutions willing to obtain the ECTS label. Then, the English version should be added to the Diploma Supplements.[7] Finally – from a general point of view – communicating in English gives universities the possibility to address to international students and teaching staff, thus raising their profile (Depraetere et al., 2011).

The characteristics mentioned in the present and the previous section (Sect. 2.1) make course catalogues an ideal test bed for the development of tools supporting translation and terminology harmonisation in the institutional academic domain. Indeed, higher education institutions would benefit from the development of such tools, since faculties have to translate roughly 2,000 CUDs on average, thus making a human translation unfeasible (ibid.). The automatisation of the translation process has been on the agenda of universities across Europe for several years now, as testified, e.g., by previous work in this area funded by the European Commission, i.e. Bologna Translation Service[8] (ibid.) and TraMOOC (Castilho et al., 2017a).[9] The former was funded in 2011 and involved ca. 10 European universities and private companies. Its goal was to train an MT system to trans-

[2]Interactive Terminology for Europe (`https://iate.europa.eu/home`

[3]More information on the Eurydice project can be found here: `https://bit.ly/2ED17Om`.

[4]`https://bit.ly/2rRVyZy`

[5]Examples of course unit descriptions can be found in the list of course units taught at the University of Bologna: `https://bit.ly/2nGbnD8`.

[6]Examples of degree programme descriptions can be found in the list of degree programmes offered by the University of Bologna: `https://bit.ly/2OEgbEh`.

[7]A model developed by the EC "to improve the international 'transparency' and fair academic and professional recognition of qualifications". `https://bit.ly/35cqtB7`.

[8]`http://www.bologna-translation.eu/`

[9]Translation for Massive Open Online Course: `http://tramooc.eu/`.

late from 8 languages – Chinese, Dutch, Finnish, French, German, Portuguese, Spanish, and Turkish – into English. Despite its interest, this project does not seem to have undergone substantial development after 2013, nor does it seem to have had the desired impact on the community of stakeholders. In addition to that, it does not include one of the language combinations examined here, i.e. Italian–English. TraMOOC – launched in 2015 – does not focus on academic courses, but rather aims at "developing high-quality translation of all types of text genre included in MOOCs (e.g. assignments, tests, presentations, lecture subtitles, blog text) from English into eleven European and BRIC languages".[9]

However, the translation of course catalogues poses a number of challenges. First, describing a course unit requires a good knowledge of institutional academic terminology, but also of expressions related to the subject matter of each unit. On the other hand – and this is particularly true for CUDs –, since they are written by non-native and non-professional translators/writers – usually teachers of the specific subject – (Fernandez Costales, 2012), their disciplinary terminology is likely to be accurate, but they might not comply with the standards of institutional academic communication. Being translated by non-professionals, moreover, often causes a decrease in the quality of their English version, thus making them unsuitable to build bilingual corpora. Finally, the English versions of the page composing a course catalogue is often much shorter than the source one. Indeed, the scarcity of high-quality bilingual texts in this domain is arguably a major bottleneck for the implementation of MT systems.

After having underlined the importance and introduced the challenges of creating an MT system to translate course catalogues, in the next sections the two text typologies composing them are described in detail, i.e. DPDs and CUDs. A comparison between the two kinds of text follows.

2.2.1 Degree programme descriptions

To the best of my knowledge, while previous work have focused on academic institutional communication in general (Callahan and Herring, 2012; Fernandez Costales, 2012) or on CUDs (Ferraresi, 2017, and references therein), DPDs or their translation have been neglected so far.

According to the ECTS Users' Guide, these texts should contain a higher number of information than CUDs.[10] The guide also introduces some of the most important contents of a DPD: the qualification awarded, the programme duration and ECTS number. The programme profile should then define the main focus of the degree programme, and the main learning, teaching and assessment activities and procedures. Also, information on the learning outcomes and on the occupational profile of graduates must be provided. For programmes with restricted access, details on the selection criteria have to be published.

As introduced in Sect. 2.2, programme descriptions present concepts pertaining both to the academic domain – e.g. ECTS, assessment methods –, and to the disciplinary field, since a description of the occupational profile or of the focus of the programme must include domain-specific terminology. However, the disciplinary contents of a degree programme are usually presented in a shallow and descriptive way, which is different from what is typical of CUDs. As a matter of fact, degree programmes address potential incoming students and thus include features that are typical of promotional texts. CUDs, being addressed to enrolled students, are more focused on contents than on style. In the

[10] `https://bit.ly/2ngHclW`

next section, after outlining the structure of CUDs, similarities and differences between these two kinds of text are introduced.

2.2.2 Course unit descriptions

Differently from DPDs, CUDs are usually composed of short sentences and of sections organised in a repetitive structure. Since they describe single units, their content is usually more specific and specialised.

CUDs are perhaps the text genre that most represents the challenges of institutional academic texts discussed so far. According to Ferraresi (ibid.), such texts were investigated in corpus and applied linguistics studies, besides being analysed for their lexical and grammatical characteristics. However, only a few works have focused on two fundamental aspects of CUDs that are related to each other, i.e. their being written in non-native English by non-professional writers/translators and their terminology (see Sect. 2.2).

The ECTS Users' Guide[10] provides a list of the elements a CUD should contain. This includes the code and title of the course, the year of study, cycle and semester it belongs to, the nature of the unit – i.e. if it is compulsory or not –, the number of ECTS credits and name of the lecturer. After this, brief descriptions of the learning outcomes, delivery mode (distance learning or lessons in classroom) and prerequisites. The following sections outline the course contents, recommended readings, teaching and assessment methods. Lastly, details are provided regarding the language in which the course is taught.

2.2.3 A comparison between course unit and degree programme descriptions

Translating – both manually or with an MT system – CUDs and DPDs poses slightly different challenges. If on the one hand both texts contain terminology and expressions belonging to the disciplinary and institutional academic domain – e.g. the ECTS credit number, the assessment, learning and teaching methods –, the contents of a whole degree programme are introduced in a more general and simple way. This is probably due to the fact that such texts are more likely to be read also by those who do not master the concepts of the specific disciplines. Being more oriented to enrolled students, CUDs offer in-depth information on the contents of the modules. Moreover, CUDs are characterised by short sentences and a clear and repetitive structure, while degree programmes usually contain longer sentences distributed in a longer text where information are not always divided in several different sections.

Turning to the challenges of applying MT to institutional academic texts, both kinds of texts are seldom (entirely) translated (see Sect. 2.2), thus lowering the number of resources available to tailor an MT system to the institutional academic domain. However, collecting an acceptable amount of bilingual course catalogues would allow to train an MT engine, working on the two research questions introduced in Chapter 1, in an attempt to streamline the expensive translation process of course catalogues also Depraetere et al. (2011) dealt with, thus helping universities to comply with the standards set up by the EHEA and to move on from the lack of terminology harmonisation institutional academic texts suffer from (Crosier, Purser, and Smidt, 2007, p.20). This process would be beneficial for the different addressees of institutional academic communication as well, since a better harmonisation and standardisation would bring an increase in quality and clarity.

Chapter 3

Machine translation: review of the literature

3.1 Introduction

In the last decade, machine translation (MT) has become one of the most used translation technologies. The 2018 issue of the Language Industry Report observed that, for the first time, more than half of the respondents had claimed to use MT.[1] This is particularly significant considering that the survey participants were translation, interpreting or localisation professionals. Moreover, the number of companies and professionals not using MT had decreased by 31% and 38% respectively compared to the previous year.

Similarly, in a survey carried out in 2018 for the 21st annual conference of the European Association for Machine Translation (EAMT)[2], 37% of the respondents answered positively when asked if they were using an MT system.[3] Also, 57% of them agreed or strongly agreed with the statement "MT helps to improve productivity". In the same survey carried out one year later, 62% of the translation companies declared to be planning investments on MT and 51% to be willing to increase its use.

Such results, together with the successful development of several neural architectures for MT (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017b) bolstered optimism in MT development. However, the whole MT history has been characterised by ups and downs, with positive results welcomed with great excitement often leading to disappointment (Castilho et al., 2018). In this chapter MT most important developments from its oldest architectures to its most recent advancements are reviewed, with a focus on their ability to handle texts from a lexical point of view, which is an issue of considerable relevance for this work and for (neural) MT usability. After a review of MT history, two of the main approaches used so far are described, i.e. phrase-based MT (PBMT) and NMT.

[1] A report by the European Commission on Expectations and Concerns of the European Language Industry: https://goo.gl/Ph4VbT.

[2] http://www.eamt.org/

[3] Survey conducted among translation professionals in 2018 first quarter: https://goo.gl/f3pXXJ.

3.2 MT history

3.2.1 MT beginnings: rule-based approaches

The year that is usually referred to as the year of birth of MT is 1949, when Warren Weaver of the Rockefeller Foundation wrote a Memorandum to outline which techniques and methods could be applied to MT (Hutchins, 1995). In the following years many events generated hype surrounding MT. The report on the state-of-the-art (SOTA) MT technologies written by Yehoshua Bar-Hillel from the Massachusetts Institute of Technology (MIT) was published in 1951. Also, the first MT conference was organised and at Georgetown University a demonstration was given in which a small batch of Russian sentences were translated into English leveraging a small bilingual vocabulary,

Three main approaches developed in the first years (Hutchins, 2007), usually classified as *rule-based* (RBMT) (see Fig. 3.1 below). The first one is the *direct* approach (Somers, 1992), in which words in the source language are replaced by their equivalents in the target one. Clearly, this approach struggles producing fluent sentences or when it comes to handling ambiguous words or sentences. The second approach is called *transfer*. In this case, the source sentence morphology and syntax are first analysed and disambiguated, then the results of the first step are transferred to the target language and finally the target sentence is reordered according to the syntactic rules of the target language (ibid.). One of the major drawbacks for *transfer* and *direct* MT systems is that they are not scalable, i.e. a new system has to be developed for each language direction.

The *interlingua* approach was based on the transformation of the source text into an abstract representation, from which the text of the target language is then generated. However, developing an artificial language that is independent from the source and target languages and able to represent a sentence abstractly was an extremely complex and expensive task. This approach would have thus ideally solved the scalability issues of the *direct* and *transfer* ones, but its development is practically unfeasible.

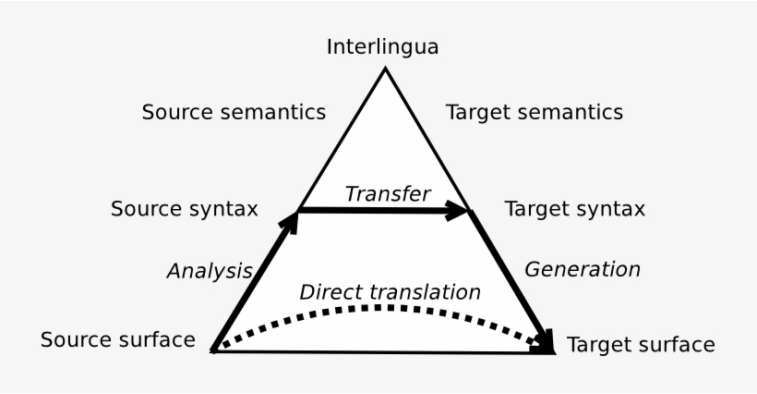


Figure 3.1: Vauqois triangle representing the RBMT approaches. The whole left side of the triangle refers to the analysis phase, while the right side represents the generation phase. Near the triangle base, semantic information is not taken into account. Near the peak, the transfer happens on a semantic level.[5]

[5]Picture taken from: `https://bit.ly/2pImUCx`.

3.2.2 Years of stagnation in MT research

Given the scalability and feasibility limits in the development of rule-based approaches, as well as the overoptimistic goal of a fully automatic high-quality translation, in 1964 the USA Government raised concerns about the actual possibility of achieving the aims for which MT research projects were funded (Arnold et al., 1994; Hutchins, 1995). At this stage, even researchers admitted to the existence of a *semantic barrier* that MT technologies were still not able to lower (Yngve, 1964). The ALPAC report (1966) maintained that the development of a useful MT system was not forthcoming yet, suggesting that sponsors focused on automatising basic translator's aids, e.g. dictionaries (Arnold et al., 1994; Somers, 1992).

Even if most of the MT research groups stopped their activities, in the 1970s the first commercial systems started to be developed (Koehn, 2010). The Météo system – developed at Montréal University – focused on the sub-language of weather forecasts for English–French translations with good results. Systran, which is still one of the most used MT systems today, was first installed to translate between Russian and English using a *direct approach*, and then extended to new language combinations such as French–English, English–Italian and German–English for the Commission of the European Communities (Arnold et al., 1994).

3.2.3 MT new era: data-driven approaches

Between the end of the 70s and the beginning of the 80s, the spread of microcomputers made it possible to develop cheaper MT systems. Research thus continued also throughout the 80s, focusing especially on interlingua systems. However, the turning point in the history of MT took place between the end of 1980 and the beginning of 1990, when after some experiments *data-driven* MT started to challenge RBMT. In *data-driven* approaches, linguistic rules are replaced by algorithms that analyse parallel corpora, i.e. collections of bilingual texts where each source sentence is aligned to its target version.

Data-driven approaches include example-based machine translation (EBMT) (Nagao, 1984) and statistical machine translation (SMT) (Koehn, 2010). EBMT looks for matches between the sentence that has to be translated and sentences in the parallel corpora. In a few steps, the input sentence is split into fragments and similarity is computed between these fragments and those in the training corpora. Similarity is mainly based on morphosyntactic and semantic information, e.g. information obtained from POS tagging and parsing.

SMT is based on algorithms that create statistical models based on the sentences observed in the provided parallel corpora and leverages these models to translate between two languages. The main statistical approach is the phrase-based one (PBMT), in which input sentences are split into phrases, i.e. contiguous sequences of words, that are then translated in the target language based on the data contained in the statistical models created during training (see Sect. 3.3.2). Other SMT architectures include the *word-based* one, where single words are the core translation unit, and the *syntax-based* one, where the fundamental units are syntactic units, i.e. parts of a syntactic tree.

IBM was one of the pioneers of this new method, and the good-quality output provided gave new impetus to research in the MT field (Hutchins, 1995). In a relatively short span of time, statistical phrase-based models (Koehn, 2010) became the predominant SMT approach (see Sect. 3.3.2 for a more in-depth description of PBMT). Also, in the very same

years the improved computational power encouraged the use of MT systems – especially in government services, companies and in the field of software localisation.

3.2.4 The paradigm shift: from statistical MT to neural MT

After several years of SMT development and use, 2015 marked a new turning point in MT history. NMT started to fill the gap with SMT (Bojar et al., 2016), eventually overtaking it even in language combinations in which SMT had long been developed, e.g. English–German (Luong and Manning, 2015).

As happened at other stages of MT history (see Sect. 3.1), this new emerging paradigm has generated great enthusiasm, with several works claiming that the gap with human translation was about to be filled (Wu et al., 2016). However, comparisons between SMT and NMT outputs for different language combinations and domains have been carried out, showing that NMT has undoubtedly brought quality improvements, but some long-term MT issues are still to be solved (Bentivogli et al., 2016, 2018; Toral and Sánchez-Cartagena, 2017).

3.3 MT architectures

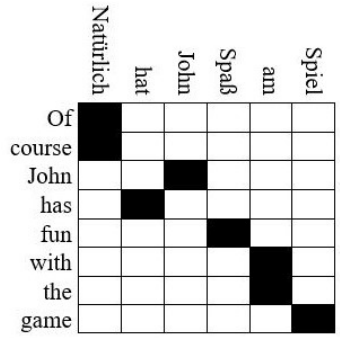
3.3.1 Introduction to MT architectures

After the historical background on MT (see Sect. 3.2), in this section the two most used MT architectures – PBMT and NMT – are described. Besides being the most used in the translation world, these two architectures are used in the present thesis as well. The main tests were carried out with NMT (see Chapter 5), but preliminary studies using PBMT were conducted as well (see Sect. 5.1). Both architectures are *data-driven* and have to be trained on a large quantity of bilingual sentences. However, they differ in the way in which they leverage these data. On the one hand, PBMT (see Sect. 3.3.2) is based on statistically-motivated phrase correspondences between the source and the target language (Koehn, 2010). On the other hand, NMT (see Sect. 3.3.3) encodes strings into vectors carrying semantic meaning, that are in turn translated into sentences in the target language (Toral and Sánchez-Cartagena, 2017).

3.3.2 Phrase-based machine translation

PBMT has been the dominant approach in the field of MT from the first years of the 21st century until 2015. Its main components are the translation model, the language model and the reordering model for the training phase and the decoder for the translation step. Bilingual sentences extracted from parallel corpora (see Sect. 3.2.3) are aligned at word-level using the expectation maximisation algorithm (Koehn, 2010), which estimates the likelihood of the alignment between a source and a target word and builds an alignment model with this information. As shown in picture 3.2, one source word can be aligned to more than one target word. The phrase extraction algorithm then extracts bilingual phrases that are consistent with the word alignments. For example, the English phrase in Fig. 3.2 “Of course John” cannot be extracted and aligned to its output sequence since its German equivalent “Natürlich John” is interrupted by the auxiliary verb “hat”. On the other hand, the English phrase “Of course John has”, can be aligned with “Natürlich hat John”. Two more examples of phrases that are not consistent with the word alignment are “has fun”

and “hat Spaß” (since in the German sentence the name “John” occurs in the middle of the phrase) and “fun with”, that cannot be aligned to “Spaß am” because it would leave out one of the words “am” is aligned to, i.e. “the”. It has to be noted that in PBMT the extracted phrases can be of different lengths, as can be seen by the phrase examples under the alignment table. Also, phrases are not intended as linguistically motivated series of words, but just as sequences of contiguous words (see Fig. 3.2).



Of course – Natürlich

Of course John has – Natürlich hat John

John – John

John has – hat John

John has fun – hat John Spaß

John has fun with the - hat John Spaß am

John has fun with the game - hat John Spaß am Spiel

Has – hat

Fun – Spaß

Fun with the – Spaß am

Figure 3.2: Example of a word-aligned sentence and sample of phrase pairs extracted consistently with the word-alignment.

The bilingual phrases extracted from each sentence pair are looked for in all training data. Based on their frequency, the probability of the target phrase being the translation of its aligned target source is computed. Each phrase pair is stored in the phrase table together with its probability. This is how the translation model, i.e. the model that includes the data needed to translate from one language to the other, is created.

The translation model alone does not ensure that the most likely phrase pairs retrieved to translate one sentence are likely to appear one after the other in the target language. This is where the language model comes into play. To build the language model, monolingual corpora of the target language are needed (usually the target side of the bilingual training data, and other large monolingual corpora when available). Sentences are split into n-grams, i.e. sequences of words of a given length (language models are usually based on 3-grams). Based on how often an n-gram occurs in the monolingual training data, the n-gram probability is computed as the probability that a word c is preceded by words a and b. The n-grams, together with their probability, are stored in the language model, which serves the purpose of enabling a fluent output by helping to choose the most likely translations and to place them in a correct and fluent word order (Koehn, 2010).

PBMT also implements a reordering model. Based on the word-alignment of the training data, the reordering model learns to predict the position of each target phrase

with respect to their source equivalent. For example, from the sentence in Fig. 3.2, the reordering model would learn that the translations of "Of course" and "John" do not occur one after the other in German ("Natürlich" and "John" interrupted by "hat").

The engine is then usually tuned on an additional small data set. The tuning technique usually applied to PBMT engines is the Minimum Error Rate Training (MERT), during which the statistical models are adjusted and adapted to a specific domain (Och, 2003).

When translating an input sentence with a PBMT system, the decoding algorithm (or decoder) leverages the models described above to retrieve and combine hypotheses, i.e. translations of a portion of the source sentence, in order to obtain the best possible translation of the whole sentence. First, an empty hypothesis is generated, which is then expanded by adding all the available translation hypotheses to translate the sentence. Typically, the choice is between a dramatically high number of different hypotheses. The decoder has to pick the right translation options and arrange them in the correct order. This search problem is typically solved by heuristic beam search. In particular, two mechanisms are typically used to reduce the search space: hypothesis recombination and pruning (Koehn, 2010).

3.3.3 Neural machine translation

The rapid succession of PBMT and NMT as SOTA MT systems has been described in Sect. 3.2.4. After describing PBMT in the previous section (see Sect. 3.3.2), in this section NMT is reviewed, starting from how neural networks work and then describing how different NMT architectures work.

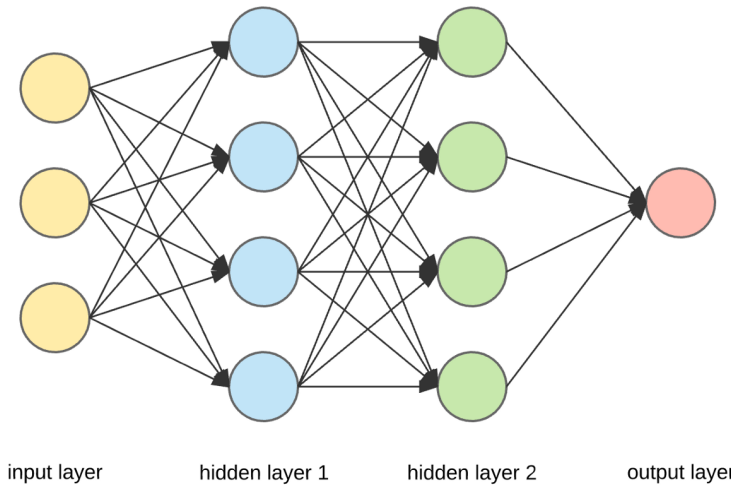


Figure 3.3: Simplified representation of a neural network with one input layer, two hidden layers and an output layer containing different numbers of nodes.[7]

Neural networks can be trained to take in an input and, through a series of operations, turn it into the desired output. They are composed of nodes structured into layers (typically an input layer, an output layer and a number of hidden layers between these two, see Fig. 3.3). Each node is connected to all nodes in the previous layer (except for nodes in the input layer) and all nodes in the following layer (except for nodes in the output layer). Nodes take in a numeric input and output a numeric output, which is then passed to the

[7]Picture taken from: `https://bit.ly/2Dnd1jb`.

nodes in the following layers. When going from one layer to the following one, these numeric values are recomputed based on the different weights and biases governing the connections between them. To learn how to produce the expected output, neural networks have to learn the best biases and weights during training.

To train a neural network, a data set including the input paired to its expected output is used. Training starts by inputting the first training item with random weights and biases. The error is then computed by comparing the output produced by the network to the output available in the training data, and weights are adjusted to reduce the error. Training ends when, after a number of iterations, the error is minimised.

More precisely, after initialising the training with random weights and biases, these are iteratively adjusted through gradient descent and backpropagation. First, the output error and the output weights are computed. The output weights are the values output by the output layer, while the output error is the difference between the produced output and the expected one. The gradient descent algorithm computes how much the output weights should be changed to reduce the output error and produce the correct output. The gradient descent is then backpropagated by the backpropagation algorithm, which starts from the output layer and moves backwards through the whole network, adjusting all weights so that the neural network becomes able to reward the output node producing the correct output (and the whole path that brought to it) and to penalise those producing the wrong outputs. To speed up the whole process, instead of computing the gradient descent after going through the whole training data set, this is randomly divided in mini-batches and gradient descent and backpropagation are performed for each of them.

Neural models applied to translation tasks are usually referred to as sequence-to-sequence models (seq2seq) (Bahdanau et al., 2014; Luong and Manning, 2015; Luong et al., 2015), i.e. models that take in an input in the form of a sequence and output another sequence. The input sequence is the source sentence and the output sequence the target one. In neural networks the input sequence is represented through word embeddings. With word embeddings, words are mapped to vectors of real numbers that have shown to capture the semantic and syntactic properties of a word. Each vector is represented in a multi-dimensional space where semantically related words are clustered together. A common example is the one where *king* and *queen* are close to each other in the vector space and the distance between them is the same that occurs between *man* and *woman*.

NMT is typically able to efficiently process a vocabulary of about 30.000-50.000 items, but the number of words necessary to translate between two languages is substantially larger. Its first implementations struggled with rare words (Sennrich et al., 2016) or with out-of-vocabulary words (OOVs). To overcome such issues these are split into subword units of variable length using the byte-pair encoding (BPE) algorithm (ibid.). Their vector representation is then computed on the subword level, which allowed to increase vocabulary coverage in NMT.

NMT engines consist of two main neural networks, the *encoder* and the *decoder* (Cho et al., 2014; Sutskever et al., 2014). The former takes in a source sequence and turns it into a vector of predefined length. The vector representing the source sentence is then passed to the decoder, which – starting from the vector – produces one output word at a time until the end of the sentence.

Different types of neural networks are used for NMT – e.g. feed-forward, convolutional – but one of the most important is the recurrent neural network (RNN) (Bahdanau et al., 2014; Sutskever et al., 2014). RNNs are particularly suitable for MT tasks because their structure allows the output of a step to be used as input for the next step. In practical

terms this gives the possibility to consider information on the previous words when producing a new one. More specifically, at a particular time step, the encoder takes in both the vector of an input word and the context vector including information on the previous words, and it updates the context vector with the information collected during this time step. This constantly updated vector is then used as input in the following time steps together with the vector of the following source items. When encoding is over, the context vector is passed to the decoder. At each decoding time step, the context vector is input into the decoder together with information on the previously produced word(s).

However, RNNs struggled with long-term dependencies, i.e. in cases where the translation of a word depends on the translation of another word but they are far away from each other in a sentence, e.g. the pronoun *her* referring to *daughter* in the following sentence: "His daughter was not feeling well and we decided to call her". Having to compress information on each word in one context vector of predefined length increases the possibility that relevant information get lost during the process. This issue was partially solved through the use of long short term memories (LSTMs) (Sutskever et al., 2014), a particular kind of RNNs that, through a series of gates, allow for deleting unnecessary information and adding relevant ones.

A substantial improvement to NMT – especially with respect to long-term dependencies – was the implementation of the attention mechanism (Bahdanau et al., 2014) to the seq2seq models described so far. The attention model is an additional neural network trained with the rest of the model. At each time step, it supports the decoder by providing information on the parts of the sentence that are more relevant to translate a particular word.

Building upon the attention mechanism, the SOTA NMT architecture at time of writing is the Transformer (Vaswani et al., 2017a), which improved the output quality and shortened the training time. The Transformer architecture is also the one used in the present work.

The Transformer is composed of a stack of encoders and a stack of decoders. Each stack is composed of a self-attention layer and a feed forward neural network. When processing a word, the self-attention layer analyses its relationship with the other words in the sentence, to encode this information in the vector. Additionally, each decoder stack implements the "conventional" attention mechanism (Bahdanau et al., 2014). This combination of different attention mechanisms leads to an improved ability of handling long-term dependencies.

Differently from previous neural architectures, the Transformer is not recursive. Just like in previous RNN models, each of the words composing a sequence is represented by a vector. However, in the Transformer, vectors flow through the encoder and decoder stack simultaneously, which means that they can be computed in parallel. This contributes to a higher efficiency with respect to the models based on RNNs, where one word only is processed at each time step.

3.4 MT quality evaluation

In MT, just like in translation in general, evaluating the quality of the target text is of utter importance. Quality is usually evaluated either manually by linguists being proficient in both the source and target language, or using automatic metrics that compare the MT output (*hypothesis* text) and one or more human translations (*reference* text) of the same source text. Intuitively, these two methods have different advantages and disadvantages.

While manual evaluation can be more accurate – taking care of linguistic aspects that algorithms would not be able to identify and with different granularity levels –, it is a time-demanding and expensive task and it can be influenced by subjective biases. On the other hand, while being able to handle more data objectively and in less time, automatic metrics might fail when subtle linguistic differences have to be considered. Perhaps more importantly, comparing a translation with one or more translations of the same sentence is an inherently biased procedure, since any sentence can be translated in many different ways in the same target language.

Regarding manual evaluation, a number of taxonomies have been developed to annotate different errors in the target text, e.g. the LISA QA model[8] and SAE J2450[9] among many others (Flanagan, 1994; Lommel et al., 2014; Tezcan et al., 2017; Vilar et al., 2006). These are usually based on the most frequent issues that can be found in translations – e.g. morphology, syntax, lexicon, grammar, and terminology issues – often divided in two macro-categories: adequacy and fluency (Lommel et al., 2014). The former measures the extent to which the source meaning is transferred to the target text. The latter measures the readability and the linguistic characteristics of the target text and can be measured without referring to its source.

Several automatic metrics were developed to measure quality based on different features. The most commonly used is BLEU (short for BiLingual Evaluation Understudy) (Papineni et al., 2002), which computes a modified precision[10] on 1-grams, 2-grams, 3-grams and 4-grams between the reference and an hypothesis text. The modified precision computes the proportion between the number of correct *n*-grams of a specific size (1-grams, 2-grams, 3-grams or 4-grams) and the total number of *n*-grams of the same size that were generated. Neither precision nor its modified version introduced here are able to take into account cases in which words that should have been generated were not generated. For this reason, if the hypothesis is too short with respect to the reference, a brevity penalty is applied to reduce the score. One of the main shortcomings in the use of BLEU is that its scores are difficult to interpret, i.e. it is hard to understand what a 18% score means in terms of quality. Also, all *n*-gram matches are given the same importance, but some *n*-grams occur less often than others and their match should therefore be rewarded with a higher score. Building upon BLEU, NIST (Doddington, 2002) (from the name of the Institute where it was developed, the National Institute of Standards and Technology) adds a higher reward for the match of less frequent *n*-grams. Also, brevity penalty is reduced for small differences in length between the reference and the hypothesis. Differently from BLEU, METEOR gives more importance to recall[11], lemmatises words that are not matched and looks for matches between synonyms. However, computing the METEOR score is costly from a computational point of view.

Different approaches are those based on edit–distance, e.g. WER (Zechner and Waibel, 2000), an automatic metric that computes the ratio between the number of insertions, deletions and substitutions and the number of reference words. Similarly, TER (Translation Error Rate) (Snover et al., 2006) computes the proportion between the number of insertions, deletions, substitutions and shifts and the number of reference words. HTER

[8]No official references are available for LISA, since this model has ceased to be developed in 2011.

[9]`http://www.mt-archive.info/jnl/LangInt-2001-Woyde.pdf`

[10]Precision in MT evaluation is usually referred to as the proportion between the number of reference items occurring in the hypothesis and the number of items in the hypothesis. BLEU score relies on a modified precision that is defined in the text.

[11]Recall in MT evaluation can be defined as the proportion between the number of reference items occurring in the hypothesis and the number of items in the reference.

(Human-targeted Error Rate) (ibid.) is the version of TER for cases in which the edit–distance is measured between an MT output and its post-edited version.

Character-based metrics, i.e. metrics based on character *n*-grams (Popović, 2015; Stanojević and Sima'an, 2015; Wang et al., 2016) have recently shown high correlation with human judgements. Their development was motivated by the need of taking into account cases in which, for example, reference and hypothesis words only differ for their suffix, but the lexical choice was correct. This is especially true for morphologically rich languages. Also, character based metrics have received more interest after the paradigm shift to NMT (see Sect. 3.2.4) and the use of subword units (see Sect. 3.3.3) (Lardilleux and Lepage, 2017; Way, 2018).

CharacTER (Wang et al., 2016) measures the number of minimum character edits required to turn the hypothesis into the reference divided by the number of hypothesis characters. Popović (2015) proposed the chrF (F-measure based on character *n*-grams) and chrF3 scores, which take into account both recall and precision between reference and hypothesis and adds a weight to reward recall more than precision.[12] ChrF3 is different from chrF in that recall is rewarded three times more than precision. The character-based metric used together with BLEU in the present contribution is CHARCUT (Lardilleux and Lepage, 2017). This metric looks for the longest substrings occurring in both reference and hypothesis, applying a length threshold that prevents meaningless matches between short strings. The rationale for the choice of BLEU and CharCut is provided in Sect. 5.3.2.

3.5 MT and lexical choices

3.5.1 Lexical issues: work comparing NMT and PBMT

The recent neural wave in the MT field has stimulated research work comparing this new architecture to its predecessor (PBMT). This section reports on differences between PBMT and NMT outputs in terms of lexical issues, since to the best of my knowledge terminology – which is one of the main focus in the present contribution – has never been specifically addressed so far.

In order to focus on lexicon, Bentivogli et al. (2016) computed HTER (see Sect. 3.4) at lemma-level between the output of different PBMT systems and one NMT system and their respective post-edited versions, reporting an improvement of 3.8% HTER points, which corresponds to a reduction of lexical errors of 17% for English–German. In this case, lexical errors included missing words, extra words and incorrect lexical choices. Building on the same approach, Bentivogli et al. (2018) observed a reduction of lexical errors for both English–German (-16,9%) and English–French (-27.1%). The analysis by Toral and Sánchez-Cartagena (2017) took into account nine language directions – between English and Czech, German, Romanian and Russian in both directions and from English into Finnish – and analysed PBMT and NMT outputs using a combination of automatic metrics. When considering the number of wrong lexical choices, omissions and extra words, PBMT performed better than NMT for two language combinations (from English into Romanian and from Russian into English), while NMT brought relatively small improvements for the other language combinations, ranging from 0.09% to 4.91%. In a study by Castilho et al. (2018), an error annotation task proved the similarity between

[12] F-measure (or f-score) is computed as the harmonic mean of precision and recall.

NMT and PBMT performance for 4 language combinations (from English into German, Greek, Portuguese and Russian) in terms of additions, mistranslations and omissions. The average number of errors showed that the NMT output contained a lower number of sentences with addition and omission issues, while the number of mistranslation errors was similar for the two architectures. However, when focusing on sentences longer than 20 words, NMT performed better than PBMT in only three cases, i.e. mistranslations for En–De, additions and omissions for En–Ru. Van Brussel et al. (2018) compared the performance of commercial systems when translating from English into Dutch, showing that the number of wrong lexical choices – including function words – is higher for NMT (304) than for PBMT (181). Interestingly, this result is even more evident when function words are excluded. NMT made 226 wrong lexical choices for content words, compared to the 91 wrong choices made by PBMT.

This overview shows that the quality leap brought by NMT may be sometimes questionable when it comes to lexical choices. If on the one hand it has to be considered that NMT has been developed for a shorter time with respect to its predecessor, on the other hand NMT complex architecture makes it difficult to find an efficient way to integrate external knowledge, e.g. domain-specific terminology, in an attempt to reduce lexical issues. This is a major concern for the industry (Way, 2018). Translation companies often have to carry out tasks in a short-time and considering customer specific requests concerning the language – and the terms – to be used. In the next two sections, work dealing with the integration of external knowledge in PBMT and NMT are presented, as a background on MT and terminology is relevant for the tests described in Sect. 5.1 and in Sect. 5.4.

3.5.2 PBMT and terminology

A number of approaches have been developed to use in-domain terminology and multiword expressions (MWEs) in PBMT, to tackle the so-called *domain adaptation* challenge, i.e. the ability of a system to adapt to a domain different from the one it was trained on. The work by Štajner et al. (2016) showed that an English–Portuguese PBMT system in the IT domain achieved best results when trained on a large generic corpus and in-domain terminology.

Langlais (2002) showed that adding terminology to the phrase table actually improved WER (see Sect. 3.4) for the French–English combination in the military domain. For the same language combination, Bouamor et al. (2012) used pairs of MWEs extracted from the Europarl corpus as one of the training resources, but only observed a gain of 0.3% BLEU points. Ren et al. (2009) automatically extracted domain-specific MWEs from the training corpora and added them to a dedicated phrase table, showing encouraging improvements in terms of BLEU score for translations from English to Chinese in the patent domain.

A sophisticated approach is the one described in Pinnis and Skadinš (2012), where terms and named entities are extracted from in-domain corpora and then used as seeds to crawl the web and collect a comparable corpus from which more terms are extracted and then added to the training data. This method shows an improvement of up to 24.1% BLEU points for the English–Latvian combination in the automotive domain.

However, all methods mentioned so far consist in adding terms to the training data, which also means that the engine has to be stopped everytime new terms have to be added to the training resources, which is a time-consuming task. Approaches to dynamically insert terminology into a PBMT system were thus investigated. The widely-used approach

in this field is the XML markup approach also adopted by Moses (Koehn et al., 2007), in which external knowledge is passed to the decoder through XML tags to force the translation of source sentence spans. However, this replaces terms in the target sentence without taking into account the surrounding words. Different solutions were thus investigated to overcome this issue. For example, Arcan et al. (2014a) tested for the first time the cache-based method (Bertoldi et al., 2013), which consists in adding to the structure of a PBMT system a dynamic language model and a dynamic translation model. The two dynamic models can be constantly updated with new automatically extracted terms. This brought about an improvement of up to 15% BLEU score points for English–Italian in medical and IT domains. For the same domains and with the same languages (in both directions), Arcan et al. (2014b) developed an architecture to identify terminology in a source text and translate it using Wikipedia. Two methods are then compared to inject terminology into the PBMT system. One is the XML markup approach described above. The other one is the Fill-up model (Bisazza et al., 2011), which consists in adding a small in-domain phrase table to the larger generic one and rewarding terms that occur in both tables, while penalising those occurring in the generic table only. Encouraging improvements – in some cases significant – over the baseline are observed on different data sets and for both language directions, while the Fill-up model consistently outperforms the XML markup approach.

Regardless of their ease of practical implementation, these approaches show that integrating terminology is possible and most of the time also successful in terms of final quality of the output. This is because the PBMT architecture, despite the complexity of its model combination, allows for intervention at different stages of its pipeline. The same does not hold true for NMT (see Sect. 3.5.3).

3.5.3 NMT, terminology and external knowledge

As explained in Chatterjee et al. (2017) integrating terminology as external knowledge into NMT systems is not a straightforward process. There are mainly two reasons for this. First, during the decoding process the NMT engine has no direct information on the position of the source word it is generating. Moreover, during the generation of a single target word in NMT, different parts of the source sentence might be considered (ibid.).

However, since the correct translation of domain-specific terminology is a central issue in (machine) translation, some approaches have been developed for the integration of terminology – or of external knowledge influencing the translation of terms – into an NMT system. Similarly to what was described in Sect. 3.5.2, methods to integrate external knowledge into NMT systems can be divided into *static* and *adaptive* (Farajian et al., 2018). In the first case it is necessary to have simultaneous access to all the training data and the information regarding their topic. An example is the work by Arthur et al. (2016), where a bilingual lexicon is exploited to compute lexicon probabilities that are then used to influence the decoder during generation of the next target word. This approach is tested on the English–Japanese language combination and improvements between 2.0 and 2.3 BLEU points are reported. Regarding external knowledge static integration, in the work by Pu et al. (2018), ambiguous words in the training data are disambiguated exploiting WordNet. A sense vector is created starting from this information and then concatenated to the word vector. However, word disambiguation might not be helpful when ability to correctly handle out-of-vocabulary (OOV) or rare words is paramount, as in the case of translation of course catalogues. Sennrich and Haddow (2016) introduced linguistic

features – e.g. part-of-speech tags, lemmas and morphological features – in their neural models to translate between English and German and from Romanian into English. Improvements ranging from 0.6 and 1.5 BLEU points are achieved adding linguistic information, computing one vector for each of them and then summing the vectors. Also in this case, features have to be available before training.

Adaptive approaches usually refer to engines trained on generic data – or on data belonging to one domain – and then tailored to a specific domain using external resources. In Chatterjee et al. (2017), the external resource is a bilingual terminology list. The decoder is extended in order to be able to consider the suggestions coming from the list of terms – provided as XML annotations – when available, and to put them in the correct place in the target sentence. This is also applicable to words that are not included in the model vocabulary. Results show an improvement of up to 3 BLEU points with respect to the baseline in two different tasks for German–English translation. Arcan and Buitelaar (2017) compared the performance of PBMT and NMT models on the translation of domain-specific expressions using different adaptation techniques. On the one hand, generic models are re-trained – or tuned, in the case of PBMT – on a small development set. On the other hand, external knowledge – i.e. terminology – is injected in the PBMT engine through XML markup (see Sect. 3.5.3), while the NMT engine looks up this information when running into an unknown word. The different techniques show that both domain adaptation and terminology injection bring better results in NMT than PBMT. However, this work only focuses on the translation of domain-specific expressions in isolation.

Differently from the approaches described so far, the one introduced in Farajian et al. (2017) is able to leverage both a large quantity of generic data and a smaller quantity of in-domain ones, achieving good performance on the translation of domain-specific terms. For the purpose of the present thesis, the ability of the system to adapt to different domains is especially key, since different domains might be present in a single course catalogue. Being able to tailor a large model on specific domains is also important to overcome the difficulties in the integration of domain-specific terms mentioned in Sect. 3.5.2.

3.5.4 Terminology evaluation

Independently from the choice of integrating terminology into an NMT engine, the ability to evaluate terminology translation quality is of the essence for MT, especially given the lexical issues discussed above (see Sect. 3.5). However, if on the one hand a number of monolingual annotated data sets for benchmarking terminology extraction and classification techniques have been created along the years for different domains (Astrakhantsev et al., 2015; Bernier-Colborne and Drouin, 2014; Kim et al., 2003; Q. Zadeh and Handschuh, 2014), the situation is much less favourable for terminology translation evaluation. Indeed, the majority of works addressing domain adaptation for MT evaluate systems only in terms of overall performance on a domain-specific test set, while very few studies specifically focus on the engines' ability to translate domain-specific terminology, and thus resort to test sets in which terms are annotated.

To the best of my knowledge, only the following manually annotated resources are made available to the community. The BitterCorpus[13] (Arcan et al., 2014a) is a collection of parallel English–Italian documents in the information technology domain in which technical terms in both the source and target sides of the bi-texts are manually marked

[13] https://ict.fbk.eu/bittercorpus/

and aligned. TermTraGS[14] (Farajian et al., 2018) is a sentence-aligned version of the BitterCorpus, which also includes a large training set.

3.6 Conclusion

In this chapter the history of MT was briefly summarised – with its ups and downs – from its very beginning until the recently developed neural models (see Sect. 3.2). The different MT architectures were described (see Sect. 3.3), with a special focus on the two dominant ones from the last few years, i.e. PBMT (see Sect. 3.3.2) and NMT (see Sect. 3.3.3). In the latter a rather straightforward statistical approach based on phrase occurrence is counterbalanced by a complex combination of different models. On the other hand, training a seq2seq model might be a less cumbersome task, but the lack of transparency in the functioning of a neural network makes development and debugging more difficult. As a result, while in PBMT many methods were investigated and applied to inject domain-specific terminology on-the-fly at decoding time (see Sect. 3.5.2), applying similar approaches to NMT is still challenging (see Sect. 3.5.3). This is arguably a major bottleneck in the application of NMT to real-word scenarios, since several works have shown that NMT still struggles with lexical choices, rare words (Arthur et al., 2016; Luong et al., 2015) and translation in low-resource or new domains (Koehn and Knowles, 2017). In spite of that, MT evaluation often considers the broader category of lexical issues without specifically focusing on terminology (see Sect. 3.5.1). The availability of data sets with annotated terminology is thus crucial for an in-depth evaluation of the MT outputs. To conclude the Chapter, an overview on MT evaluation was provided (see Sect. 3.4), including the automatic metrics that are used in the present thesis.

[14] `https://gitlab.com/farajian/TermTraGS`

Chapter 4

The concept of trust: review of the literature

4.1 The concept of trust

People become acquainted with the concept of trust at an early age and hear this word in a large number of conversations happening every day and regarding different topics. As for many other widespread terms that can be applied to different fields, it is not easy to find a universal definition for trust (Blomqvist, 1997; Madhavan and Wiegmann, 2007). The first two definitions in the online Oxford Dictionary are: "Firm belief in the reliability, truth, or ability of someone or something" and "Acceptance of the truth of a statement without evidence or investigation".[1] McKnight and Chervany (2001) argue that in three different monolingual English dictionaries on average 17 different definitions of trust are provided. To introduce the following paragraphs, it is useful to start with a simple definition by Lee and See (2004): "Trust is the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability". The authors introduce some of the fundamental characteristics of trust, i.e. there has to be a situation of uncertainty and two parties involved, one of which is vulnerable and depends on the other one. Even if these definitions seem to exclude the presence of trust in relationships other than human-human ones, given the nature of the present work it is important to notice that trust can also occur in faceless situations, i.e. in situations in which there is no contact with other people (Glanville and Paxton, 2007). Anyway, the complexity of trust and its application to a number of domains requires to further review its definitions and some of the most important concepts related to it.

According to Luhmann (2000), for example, trust comes into play every time expectations are involved that could lead to disappointment. However, the author further specifies that such a statement could also apply to confidence, and the difference is that confidence involves low risk and a situation in which alternatives are either not available or neglected. For example, one is confident that he/she will not be assaulted while leaving the house. On the other hand, trust involves making a decision being aware of the risk(s) and of the possible alternatives. You might decide not to rely on ready-made food quality only if you are willing to spend time cooking something for dinner. Similarly, a translator can choose not to rely on an MT output only if willing to produce a translation from scratch and, in any case, the risks (decrease in productivity and/or lower quality) and alternatives (translation from scratch or not accepting the task) are well-known. Since this is the most likely

[1] https://en.oxforddictionaries.com/definition/trust

scenario in a translator's workflow, the remainder of this work focuses on trust rather than confidence.

Lee and See (2004) framed the concepts of trust and reliance following the definitions proposed by other authors. A human being has intentions that are expressed by his/her behaviours. Behaviours are influenced by attitude, which in turn is influenced by perception and beliefs. This scheme defines the relationship between trust and reliance, where trust is the attitude and reliance the behaviour. The difference between trust and reliance is crucial to this work. Translators might rely on translation memories (TM) because they trust them, while they might not rely on an MT output because they do not trust MT in general or a specific MT system. A different definition of reliance is the one by Blomqvist (1997), according to which reliance is "the trust that keeps society going [...], i.e. that something will happen". However, this general definition is not appropriate for the present work. Since the goal here is to frame professionals' behaviour when post-editing and when translating, Lee and See scheme described in the previous lines is applied to understand whether translator trainees have perceptions and beliefs that negatively influence their trust – attitude – towards MT in a way that prevents them from relying – behaviour – on the MT output.

Another way to model trust that is also applicable to the field of technologies is the one described by Madhavan and Wiegmann (2007) and Rempel et al. (2004). The first element modeling trust is *predictability*, i.e. if one individual behaves in a predictable and/or consistent way over a span of time. The second one is *dependability*, which can be defined as the extent to which a person is confident in the trustee. Also, *faith* is an internal characteristic of a person and it refers to the willingness of using an aid to carry out a task. In terms of internal characteristics, Blomqvist (1997) suggests that the concept of credibility and trustworthiness are strictly related to each other and sometimes also confused. However, when someone is credible it means that he/she has the necessary resources to perform an action he/she has claimed to be able to perform. In turn, credibility is similar to competence, i.e. if the trustee is perceived as being able to carry out a task. Thus, competence depends on how an individual is perceived independently from his/her actions or statements. Finally, Blomqvist (ibid.) suggests that a distinction must be made between hope, which does not imply unpleasant outcomes if it is not fulfilled and trust, which does. Even if concepts such as *credibility*, *dependability* or *faith* are easy to be linked to the translator-MT relationship – e.g. the credibility of the MT system established if it is able to produce a good-quality translation –, in the next sections the way in which different authors have presented such notions in order to apply them to a human-machine relationship – which is of the essence for the present work – is analysed.

4.2 Building Trust

The debate around this multifaceted notion is not only focused on the most appropriate way to define trust and how to discern it from similar concepts, but also on the way in which trust forms in a person, if it evolves, if it is influenced by external factors, etc. These issues have been investigated by several studies whose conclusions can be grouped in two different approaches.

On the one hand, as stressed in the works by Blomqvist (1997) and Lee and See (2004), some researchers see trust as an enduring personality trait or a psychological propensity (Glanville and Paxton, 2007). According to these theories, trust forms in the early years of our life and is not likely to be influenced by experiences. As a result,

there are high-trust human beings and low-trust human beings with different behaviours according to their level of trust.

In contrast to the idea of trust as an immutable trait, several studies have developed the social learning perspective (Glanville and Paxton, 2007) or social learning approach (Lee and See, 2004). One's level of trust changes not only over time, but also across different kinds of human relationships based on previous experiences. Hence, it is possible for a person to show high-trust when relating to the neighbors and to show low-trust when cooperating with a group of colleagues with which he/she had previous negative work experiences.

Lee and See (ibid.) describe a similar way in which trust develops, yet referring to it as "attitude". Reviewing the works by other authors, they argue that the initial level of trust is given by past experiences or even by gossip. Then trust changes along with the relationship. As mentioned before, the way in which Lee and See (ibid.) frame trust is the most relevant for this work. Also in the case of translators asked to work on MT outputs, the attitude at the beginning might be influenced by gossip and then change if their customer provides them with a good-quality output to work on. With respect to trust development, Blomqvist (1997) points out that, in a relationship, while the first evidences of untrustworthiness are usually seen as accidental by the trustor, the following ones fuel a vicious circle of distrust generating distrust. The same happens when the trustworthy behaviours of the trustee generate always more trust in the trustee. In the case of post-editing, while it is not possible to maintain if the first evidences of untrustworthiness are seen as accidental by the translator, it is likely that a high number of issues generate consistently more distrust towards MT and vice versa when a small number of issues is detected.

4.3 Different kinds of trust

In the previous section (see Sect. 4.2) the way in which trust is built was investigated. Anyway, trust is influenced by many factors that are not limited to the trustor disposition or his/her past experiences. In this section different kinds of trust becoming established because of factors external to the trustor are analysed.

If trust is thought of as a grammar as in McKnight and Chervany (2001), also the direct object has an influence on the different kinds of trust. Reviewing the works by many authors Blomqvist (1997) suggests that with *specific, personal* or *particularistic* trust, different authors refer to the kind of trust that establishes when the trustor knows the trustee personally. On the other hand, *generalised* trust forms in situations in which one has to rely on strangers, e.g. for people working in big companies.

So far, human-human relationships were described, but trust is also relevant in relationships where the two parties are of a different nature (Lee and See, 2004; Madhavan and Wiegmann, 2007). This is the case with *organisational* trust, which involves an organisation's identity (Blomqvist, 1997). A similar case is that of *institutional* or *system* trust, where there is no direct contact with known people and the trustor assumes that a system will operate in a predictable way: "Institution-based trust is situation-specific but cross-personal because it means that one trusts the specific situation but does so irrespective of the specific people in that situation" (McKnight and Chervany, 2001). Blomqvist (1997, and references therein) further clarifies that "*institutional* or *system* trust can serve as a substitute for the need to trust at the interpersonal level" and it does so by substituting the necessity to trust one other person. One common example is the International

Organization for Standardization (ISO)[2], which makes sure that products, companies or professionals comply with common standards before certifying them. The ISO label confirms the trustworthiness of a professional also in cases in which the trustor does not know the professional personally (see Sect. 4.4.3 for another example of system trust). Similarly, a translator working with a TM would probably choose not to rely on sentences previously translated from another unknown translator if he/she was not sure that a translation agency carry out a revision or quality check on the final work before populating a customer's TM. The same should apply to MT, i.e. a translation agency should guarantee for the quality of a raw output assigned to a translator.

This section has introduced how trust can take different forms based on external factors, e.g. the characteristics of the trustee or the fact that the trustee might be known or not. However, these different kinds of trust might further change when they are applied to different domains. Next section analyses how trusting relationships are affected by the domain in which they are established.

4.4 Trust in different domains

4.4.1 Introduction

As previously mentioned (see Sect. 4.1), one of the reasons why trust is a difficult concept to frame is that it applies to several domains in which it plays slightly different roles. In addition to psychologists, many other specialists from various fields have tried to define trust. Then, in the last few decades the continuous growth in the use of IT products and the increase in distance and faceless relationships have paved the way for a new understanding of trust. In this chapter a brief definition of trust in marketing and contract law is provided (see Sect. 4.4.2), i.e. two examples in which the concept of trust has not been deeply influenced by new technologies. Then, changes in the concept of trust for domains highly influenced by the growth in the use of technologies are considered (see Sect. 4.4.3 and 4.4.4), in order to create the necessary background to handle trust in the field of MT (see Sect. 4.4.5).

4.4.2 Trust in contract law and marketing

The role of trust in the domain of contract law is particularly interesting. Since contracts are drawn up to protect each party from any damaging behaviour of the counterpart, one might think that trust is of minor concern or even that it can be replaced by a contract. However, at the very beginning of a cooperation between two parties no contract is in force yet. When these are in force and one of the parties fail to respect them, litigation are not common since they can be slow and expensive (ibid.). For these reasons, "lawyers writing about trust tend to see it as a necessary complement to the control of formal legal contracts" (ibid.).

Blomqvist (ibid.) briefly analyses how trust influences interactions in the field of marketing. The most interesting example is the relationship between the buyer and the salesman, where the former has to be willing to trust, while the latter must prove himself/herself a trustworthy person.

[2] `https://www.iso.org/home.html`

More in general, trust in marketing can be defined as a "long-term attitude among individuals or companies" (Blomqvist, 1997) where drawbacks in a relationship are allowed as long as both parts show willingness to meet the other one's needs.

4.4.3 Trust in e-commerce

Trust in e-commerce is the development of trust in marketing (see Sect. 4.4.2) and is also one of the modern scenarios in which new technologies have brought about a new way of seeing trust (see Sect. 4.4). As a matter of fact, the fast-paced growth of e-commerce has raised the question of how to create a virtuous circle of trust generating trust (see Sect. 4.2).

McKnight and Chervany (2001) mention some of the information that vendors can share with the (potential) buyers in order to appear trustworthy. This is a way to implement the trust constructs seen so far (see Sect. 4.1). First, providing for a privacy policy or a third-party seal might be seen as a way to establish trust, since – in a faceless interaction – an ethical attitude or the approval of a third party might influence the buyer's willingness to trust. In particular, the third-party seal falls into the category of institutional trust (see Sect. 4.3), i.e. in a faceless interaction one might rely on an institution that guarantees for another party's trustworthiness. Other actions to be taken might be creating links to other websites, and reputation building (ibid.). The latter plays a key role especially because the perceived trustworthiness can also be influenced by gossip (see Sect. 4.2). As a matter of fact, a deeper, benevolence-based trust establishes only after a number of transactions – which is rarely the case in e-commerce –, so reputation and economic reasoning are of utmost importance (Ba and Pavlou, 2002).

The way in which users often judge the reputation of an online vendor is through feedback mechanisms. The work by Ba and Pavlou (ibid.) showed how enforcing the buyer's trust can generate higher incomes and bring higher price premiums especially for expensive transactions, i.e. a vendor with a high average feedback rate can increase his/her prices with respect to a vendor with a lower feedback rate selling the same product(s). The higher the transaction value, the higher the incomes.

The reason for this focus on e-commerce is the link that can be drawn between trust in this field and trust in the MT field. The post-editing of MT has brought new kinds of interactions between professional users and the texts. The shift from revising a text produced by another professional translator to the post-editing of a text produced by a machine might be viewed as similar to the shift from buying from a local store to ordering from an online store. It is therefore possible that finding ways to increase MT credibility among post-editors – e.g. through training – is of the essence in order to breach the vicious circle of distrust generating distrust and to make the cooperation between trustee and trustor advantageous. This would confirm what was maintained in Ba and Pavlou (ibid.), where the feedback mechanism was found to contribute to e-sellers' credibility increase, generating price premiums.

4.4.4 Trust in technologies

In the previous sections some of the domains to which the concept of trust applies were reviewed. In each of them (e.g. Sect. 4.4.2 and 4.4.3), however, the trustee was a person (even if, in the case of e-commerce, the trustor does not always have a direct contact with the trustee). In this section trust is looked at from a new perspective, i.e. shifting from

trust between humans to trust between a human and a machine.

The topic of trust in a human-machine interaction has become increasingly important with the dramatic growth in the use of technology aids in various activities, fields and in our everyday life. Researchers have focused on this topic arguing that – even if human-machine relationships may develop in the same way as human-human ones (Madhavan and Wiegmann, 2007) – the constructs developed to describe trust between human beings cannot hold true for trust in human-machine interactions (Lee and See, 2004). One fundamental difference is that human beings behave intentionally, which is one of the most important factors in human-human trust (ibid.). At the same time, technologies are created by human beings and are therefore biased by the developer's intentionality. Moreover, users tend to perceive machines' behaviours as intentional when their technologies are particularly sophisticated and can mimic human actions.

The second difference is that interpersonal trust is symmetric since it depends on how both parties perceive the counterpart's behaviour. This does not happen when the trustor is a machine, so the trust relationship develops in a different way (ibid.). In the field of technologies, "attributions of trust can be derived from a direct observation of system behaviour (performance), an understanding of the underlying mechanisms (process), or from the intended use of the system (purpose)" (ibid.). According to Madhavan and Wiegmann (2007) process and performance correspond to dependability and predictability, while purpose corresponds to faith (see Sect. 4.1). At the beginning, thus, the user might be aware of the intended use for a software – e.g. when detailed documentation is available – but at the same time, information on its performance might not exist. In this case, trust is based on purpose and not on performance.

The way in which trust builds obviously influences reliance in the technology used. Lee and See (2004) describe this relationship as a loop where lack of trust brings a lack of reliance and the lack of reliance causes the absence of reliable information regarding the technology and its use. When no details are available on the (correct) functioning of a technology, the initial level of trust depends on the user's predisposition and then on the observations of the technology's reliability.

User's predisposition and self-confidence affect the choice to rely on a system in many ways (Lee and See, 2004; Madhavan and Wiegmann, 2007). If a user is self-confident and has a low level of trust in automation, he/she is likely not to rely on the latter and vice versa. Self-confidence might also be combined with experience. For example, a study comparing the behaviour of students to that of pilots when using a driving aid system showed that students were less likely to rely on the technology aid. This might be caused by more limited experience with the automatic aid with respect to professional drivers (Lee and See, 2004). Similarly, Yang et al. (2017) argue that a user evaluates the usefulness of an automation based on his/her ability to carry out the task without any help. If one is capable of carrying out a task alone, technology is evaluated less favorably.

Faults are another fundamental factor in the relationship between trust and reliance when the trustee is a machine or a software. However, the way in which faults affect trust and reliance depends on several factors. The moment in which the fault happens – i.e. if the performance is very good at the beginning and then decreases, long-lasting trust is more likely to establish. If poor performances are encountered at the beginning, trust is less likely to establish, and below a certain threshold of reliability it will decrease even more rapidly (Lee and See, 2004). Moreover a small, systematic fault – that can be controlled and does not compromise the final result – does not affect trust and reliance in a decisive way (Lee and See, 2004; Yang et al., 2017). When a user notices a fault

with unpredictable consequences for his/her task, trust will decrease independently from the fault's dimensions. Yang et al. (2017) also showed in their work that when a user has positive experiences in using an automation, a positive feedback loop establishes, i.e. the more a technology proves itself trustworthy, the more the trustor will trust it and choose to rely on it. In the opposite case, a negative feedback loop establishes that has a stronger effect on trust than the positive one.

However, the extent to which faults influence reliance also depends on the user's perception. Sometimes, mismatches between trust and the machine's reliability cause a lack of reliance that is not directly caused by objective issues. To describe the balance between user's trust and machine's reliability Lee and See (2004) introduced three parameters: calibration, resolution and specificity, that are also reviewed in Madhavan and Wiegmann (2007). Calibration is the match between a user's trust and the hardware/software capabilities. When trust is higher than the capabilities, overtrust emerges. Mistrust occurs when the machine's capabilities exceed the level of trust. "Resolution refers to how precisely a judgement of trust differentiates levels of automation reliability" (ibid.), e.g. if there are various degrees to which the automation capabilities can change but only a few possible trust degrees, than resolution is low. "[...] specificity refers to the degree to which trust is associated with a particular component or aspect of the trustee" (ibid.), e.g. specificity is high when trust is strictly connected to the different capabilities of different modules of the system. When these three parameters are high, the level of reliance is appropriate for the specific technology, and misuse – excess of reliance – and disuse – failure to rely on the automation despite its capabilities – are avoided. Clearly, the way in which misuse and disuse refer to reliance is similar to the one in which overtrust and mistrust refer to calibration of trust. Appropriate reliance and calibration can then lead to a more effective training, design and evaluation in order to enhance human-machine partnership (Lee and See, 2004).

In a situation in which misuse, disuse, overtrust or mistrust are avoided, the human-machine interaction brings positive results, improving the productivity of a human alone: "Appropriate trust can lead to performance of the joint human-automation system that is superior to the performance of either the human or the automation alone" (ibid.). The way in which the cooperation between human and machine can be enhanced is particularly relevant for this work. According to Lee and See (ibid.) this goal can be reached, among other methods, by:

- Designing for appropriate trust and not for higher trust.
- Providing information regarding past performances.
- Making the algorithm and its (intermediate) processes understandable for the user.
- Training users on the technology's reliability, its intended use and the mechanism underlying its behaviours.

4.4.5 Trust in MT

Despite the evidence of its importance in various fields (see Sect. 4.4), to the best of my knowledge trust among MT users was rarely investigated, except for the two contributions described in the present section.

Recently, Cadwell et al. (2018) interviewed two groups of professional translators to investigate the reasons behind the adoption or rejection of MT suggestions. Though with

some differences – probably due to their different work environment – the two groups mentioned the lack of trust toward MT as one of the reasons for the non-use of MT segments. One of the most interesting points emerging from this work is the way in which the human factor influences and guides trust. This can be summarised in two categories: when the work by professional translators is compared to the one of an MT engine, and when human professionals can influence the trust of others towards the technology. In the former case, both groups claimed that – even though they do not trust all translators equally – they do trust a human translation more than one produced by a machine. Also, one of the groups stated that being aware of the fact that MT leverages previous translations produced by professionals slightly increases trust towards the output, whereas the other group stated that this awareness is not sufficient. As a matter of fact, while translation memories (TMs) are directly fed by translators and provide useful information about the author and the creation date of a specific entry, MT suggestions only come with information on the engine that produced them. For the second category, one participant stated that being able to directly contact the in-house MT system developer produces a slight increase in his/her trust. Even if Cadwell et al. (ibid.) stated that these results can not be generalised and applied to similar or different groups, it is interesting for the remainder of the present work to notice that participants are actually coping with a lack of trust towards MT trying to find elements that allow them to trust this technology more. This is similar to what was mentioned in paragraph 4.4.3 describing the third-party seal (McKnight and Chervany, 2001) or the feedback mechanism (Ba and Pavlou, 2002). Also, the fact that one whole group complained about the different amount of additional information provided by a TM and by an MT engine, might seem to confirm that professionals are relying on extra information provided in the post-editing environment.

Going back to practical post-editing tasks, Martindale and Carpuat (2018) chose to test the influence of good – fluent and adequate – and bad – fluent but not adequate or adequate but not fluent – translations on non-professional translators (see Sect. 3.4 for a definition of fluency and adequacy). A translation is fluent when its grammar is correct and it sounds natural in the target language, while it is accurate when the source text message is preserved in the target text (Specia et al., 2011). Each participant saw an MT output and was required to specify his/her level of trust for that translation and for the MT system in general. This operation was repeated after having seen a human translation for the same segment. Results suggest that fluency issues have a stronger negative impact on non-professional translator's adequacy errors. According to the authors, the great impact of fluent translations on the users' trust highlights the necessity for MT developers to either reduce the number of fluent but not adequate translations or to flag them as a warning for the user. Since users place great importance on fluency, they are likely not to notice adequacy errors in the target sentence (e.g. omissions or additions).

Despite its small-scale nature, the work by Martindale and Carpuat (2018) shed light on several important factors influencing trust in the non-professional use of MT systems and therefore potentially preventing a productive use of MT. Most importantly, it confirmed that the output quality is not the only factor for a fruitful human-machine interaction in MT. User's trust might be influenced by specific features in the output and their frequency. Addressing output quality issues in general terms might therefore not be enough to enhance the interaction between post-editors and MT.

Chapter 5

Assessing the feasibility of applying MT to institutional academic texts

5.1 Introduction

Bilingual communication is a key factor for the internationalisation of universities, causing the need to translate a large amount of texts into English every year. Streamlining the translation process would be beneficial for universities, especially for those based in countries where English is not an official language. Indeed, several projects aimed at enhancing the efficiency of the translation process through the use of translation technologies did not bring the desired impact (see Chapter 2), neither from a qualitative, nor from a quantitative point of view.

This lack of results is not surprising, since many factors make the production of multilingual versions of institutional academic texts particularly problematic. Lack of high-quality bilingual versions and lack of standardisation between texts produced by different higher education institutions (see Chapter 2) are some of the most relevant challenges for the development of tools handling the translation of degree programme descriptions (DPDs) and course unit descriptions (CUDs), i.e. the two main text types in course catalogues. Also, it has to be noted that what is referred to as *institutional academic domain*, is actually a multi-domain scenario, since it contains texts and terms belonging to the educational domain and to different disciplinary domains, e.g. biology, chemistry, economics. The presence of multi-domain terminology is thus a further challenge to be faced during the translation of such texts.

Nonetheless, given the urgency of developing such tools and the advances in MT brought by the use of neural networks, this chapter tests a pipeline to apply NMT to institutional academic texts. The language combinations examined are from Italian into English and from German into English. As stated in Chapter 1, this project and its results can be profitable for universities, contributing at the same time to some of the research fields that need the most attention when it comes to (N)MT, i.e. terminology translation and low-resource domains.

Training in-house customised NMT systems is not a viable option for institutional academic texts, since a high amount of in-domain sentence pairs would be required. Identifying the most suitable ready-to-use SOTA MT system(s) is thus of the essence to reach good performance when no bilingual data are available. At the same time, the possibility that even a small amount of data might become available has to be considered. For example some universities might have a small quantity of sentence pairs available, or new

sentence pairs might have been produced translating and post-editing the first course catalogues. A suitable MT system for this pipeline should provide the possibility to leverage even small amount of bilingual data when available – if this helps to improve the output quality – at the same time producing an output of good quality when no in-domain data can be exploited.

In the present Chapter, two SOTA ready-to-use NMT systems are tested: ModernMT (MMT)[1] and Google Translate (GT).[2] Both of them are based on the state-of-the-art Transformer architecture (see Sect. 3.3.3) and trained on a large pool of parallel data. If able to reach a good quality in this domain, MMT would be ideal to be integrated in the pipeline presented here, since it implements an adaption mechanism which allows the system to adapt to new data – if available – in real time (Bertoldi et al., 2018). As a SOTA system, GT provides an external validation of the quality of MMT, although it does not provide any adaptation mechanism.

The two MT systems are tested in two realistic scenarios for universities willing to use MT. In the first one, no in-domain data are available. GT and MMT are tested, the latter in its generic version. In the second scenario, good quality bilingual course catalogues become available. In this scenario, only MMT can be used, since GT cannot be adapted.

Simulating these two scenarios requires the collection of parallel data for both language combinations (It–En and De–En). Specifically, a smaller data set is needed as test set, while a larger one can be used for domain adaptation. For this purpose, sentence pairs were extracted from the websites of 4 different Italian and German universities and randomly assigned to the domain adaptation or test sets. Following the data set creation, an overall evaluation of the MT systems was carried out using automatic metrics as to provide an overview on the quality that can be achieved in both scenarios for each language combination.

As mentioned above, a high density of multi-domain terms characterises institutional academic texts. Evaluating term translation in the two scenarios would on the one hand contribute to a more fine-grained understanding of the output quality, while on the other hand it would contribute to fill the gap in the research on terminology translation for NMT (see Sect. 3.5.4). Given the lack of resources to evaluate term translation, a whole new pipeline and specific resources have to be conceived and built, which is a time-consuming task. For this reason, terminology evaluation is carried out on Italian–English only. The MAGMATic (Multi-domain Academic Gold Standard with Manual Annotation of Terminology) data set was created and used. It includes 2,055 Italian–English sentence pairs with manually annotated terms on the target side. Each term is also assigned to its specific domain. MAGMATic was presented in Scansani et al. (2019b) and is released under a Creative Commons Attribution – Non Commercial – Share Alike 4.0 International license (CC BY–NC–SA 4.0), and is freely downloadable at: `https://ict.fbk.eu/magmatic/`.

MT outputs used for the overall evaluation and the terminology assessment were produced on February 5th, 2019. To the best of my knowledge, they represent the first attempt at applying NMT to institutional academic texts.

Between the end of 2016 and the beginning of 2017, tests were also carried out on the other dominant approach at that time, i.e. PBMT (see Sect. 3.2.3 and 3.3.2). The collected bilingual texts were split into training, development and test data sets for It–En and for De–En and different engines were built using the phrase-based open-source version of

[1] `https://www.modernmt.eu/`

[2] `https://translate.google.com`

MMT (Bertoldi et al., 2017). Terminology translation was investigated adding the IATE termbase for the education domain to the training data set.[3] More in detail, Scansani et al. (2017b) focused on the impact of terminology on the overall output quality, measured in terms of BLEU score. The authors compared the performance of a baseline trained on a subset of the Europarl corpus against an in-domain engine trained on the aforementioned data set.[4] In a second step, both engines were enriched with the IATE education termbase. Results showed that the use of terminology did not improve the overall quality substantially, especially for the engines trained on the in-domain corpus. For Italian–English, the baseline engine with terminology reached a 22.75 BLEU score (22.58 without terminology), while the in-domain engine with terminology (29.82 BLEU) was slightly outperformed by the one without terminology (30.60). For German–English the baseline scored 34.98 BLEU without the termbase and 36.89 with the termbase, while the in-domain engine reached a 46.31 BLEU score without the termbase and 47.05 with the termbase. The absence of substantial improvements after the integration of terminology makes the results of this preliminary study unsatisfactory. Building upon the same method, Scansani et al. (2017a) focused on terminology translation in terms of F-measure for the language combination Italian–English and trained two engines on both the subset of the Europarl corpus and the in-domain training data set. One of the two engines was further enriched with the IATE education termbase. F-measure was computed on the number of English termbase entries appearing both in the reference and in the output. Results were 0.568 for the engine using terminology and 0.577 for the one without terminology. Even if a subsequent manual analysis showed that the termbase was sometimes able to influence the choice of the correct term, F-measure had already proven that the use of the termbase did not impact on the output quality. This is probably due to the fact that the termbase entries occurring in the test set appeared in the training corpora as well.

According to the two preliminary studies, using customised PBMT engines and terminology for institutional academic texts did not yield satisfactory results. Adding in-domain terminology to the data set did not improve the overall output quality or the ability to handle term translation substantially. Further studies could have examined the injection of terminology from other domains (e.g. the disciplinary ones), or other sources, and compared different injection technologies (see Sect. 3.5.2). However, given the multi-domain nature of institutional academic texts, choosing the right termbase to be added to the translation model and the most suitable technology to inject relevant terms would have been a complex task. Also, the simultaneous first successful applications of NMT in different domains led to the decision to set aside the complex and unpromising tasks planned for PBMT to move to the NMT tests described in the following sections and motivated above.

The next section (5.2) describes the method used to build parallel corpora and to split them into different data sets. Following that, the overall evaluations on the NMT engines and their results are described for each scenario and discussed (see Sect. 5.3). To conclude, the work that led to the creation of the MAGMATic data set and the subsequent term assessment are described (see Sect. 5.4).

[3] IATE (InterActive Terminology for Europe) is the European Union multilingual terminology dictionary: `https://iate.europa.eu/`.

[4] Europarl is a parallel corpus composed of texts produced by the European Parliament. `http://opus.nlpl.eu/Europarl.php`.

5.2 Building parallel corpora

5.2.1 Data collection methodology

At the start of the present research project, an Italian–English course unit description corpus was already available thanks to the CODE project.[5] This is composed of CUDs published on the website of the University of Bologna. Besides those included in the CODE corpus, more sentence pairs were needed to enhance the Italian–English data set and to build a German–English one from scratch.

Assuming that providing bilingual versions of web contents increases the web impact and reputation of university websites, Webometrics – a website that ranks the Higher Education Institutions of the whole world based on their web presence and impact – was used as a reference to identify, for each language combination, university websites from which bilingual texts could be collected.[6] Based on the top 20 for each of the two countries, CUDs and DPDs were looked for and chosen if they satisfied the following criteria: (*i*) good-quality bilingual contents, (*ii*) presence of a good amount of text, (*iii*) ease of extraction. Websites whose web contents were not translated or whose translations were not comparable with the source text (e.g. short summaries rather than real translations) were discarded. Websites whose CUDs and DPDs were split into several tabs containing a small amount of words each were not taken into account, since they would have slowed down the whole collection and cleaning process. The analysis of the Webometrics top 20 institutions from Germany and Italy indicated that web contents of the faculties whose discipline belonged to the macro-category of the humanities rarely met criteria *i* and *ii* above. This macro-category was thus excluded. This also helped to control the number of domains considered, allowing for a more focused and in-depth evaluation on them, especially in the terminology assessment. Also, including humanities would have introduced domains that are very distant from each other, e.g. theology and mechanical engineering, reducing the data set coherence.

The webpages and links used to build these corpora were harvested from November 2016 to March 2017 and then revised in November 2017; for this reason, it is possible that the texts are not available on the various faculty websites anymore, due to changes in the course units or degree course programmes.

5.2.2 Inspection of Italian academic websites

As stated in section 5.2.1, a CUD corpus was already available thanks to the CODE project. Since humanities were excluded from the scope of the present work, CUDs belonging to this macro-domain had to be removed from the CODE corpus, which included humanities as well. Each unit description in the CODE corpus starts with a header – composed of the course name preceded and followed by three asterisks – that made a quick identification of the disciplinary domain possible. Those belonging to humanities were then removed. Some of the disciplinary domains covered by the CODE corpus after maintenance include economics, physics, chemistry and biology.

The University of Bologna (Unibo) is also the top Italian university according to the Webometrics ranking. Moreover, the DPD section of its website complies with the criteria

[5]CODE is a project funded by the University of Bologna through the 2013-2015 FARB scheme. It was aimed at building corpora and tools to support translation of CUDs into English and drafting of these texts in English as a lingua franca. `http://code.sslmit.unibo.it/doku.php`.

[6]Webometrics link: `http://www.webometrics.info/en/Methodology`.

listed in Sect. 5.2.1. The available corpus was therefore enriched with DPDs from this university.

The Italian–English text collection was augmented with CUDs extracted from the website of the Politecnico di Torino (Polito). According to the objectives listed in section 5.2.1, the website of the Politecnico di Torino (Polito) was chosen since it contained a good amount of bilingual contents clearly split according to language and with a good amount of sentences for each page.[7] Moreover, the Italian Politecnico is a technical higher education institution, its disciplines – e.g. information technologies, physics, mechanical engineering etc. – thus rarely belong to the humanities, which complies with the criteria in Sect. 5.2.1.

Two Italian university websites for which bilingual CUDs were not available were chosen: University of Rome Tor Vergata (or UniRoma2)[8] and Politecnico di Milano (or Polimi)[9].

As previously mentioned – see Sect. 5.2.1 – some of the web pages found in the Webometrics ranking were discarded because they did not fit the requirements in Sect. 5.2.1: alignable texts, acceptable amount of sentences and ease of extraction. Examples are provided below. La Sapienza University does not provide an English version of its degree programme or of its CUDs.[10] Similarly, at the Federico II University of Napoli, bilingual descriptions are available only for degree programmes taught in English.

Another interesting example is the one of the University of Torino. On this website, many CUDs are provided both in Italian and in English.[11] However, the page is split into many tabs containing one part of the description each – i.e. one tab for the aim of the course, one tab for teaching methods, etc. Only the text inside the tabs is provided in the two languages, while the remaining contents are in Italian. Thus, examining the exact amount and quality of translated contents would be a cumbersome task. Moreover, such texts would require a time-consuming removal of non-bilingual sentences after extraction.

Some of the departments at the University of Firenze provide an English version of their DPDs, but it is just a short version of the Italian text, thus the two versions are not alignable. Another example is the website of the University of Trento, which does not contain any translated description.

5.2.3 Inspection of German academic websites

Similarly to what was observed for Italian–English, when German is the source language, the scarcity of high-quality bilingual texts is also arguably a major challenge. In addition, for the German–English language combination no corpus was already available at the beginning of the project. Therefore, the whole benchmark was created based on the approach described in Sect. 5.2.1.

Based on the Webometrics ranking and aiming to build a bilingual resource as comparable as possible to the Italian–English one, four of the top-ranked German universities were targeted. Only disciplines not belonging to the humanities area were selected (see Sect. 5.2.1). The four universities chosen are the Technische Universität München

[7] E.g. CUDs for the Biomedical Engineering Master of Science: `https://goo.gl/DWT8ck`.

[8] `https://goo.gl/kpQm4Q`

[9] The contents extracted from this website are not available anymore, due to website renovation.

[10] An example of degree programme found in the University of Roma La Sapienza website: `https://goo.gl/hF59Vi`

[11] A description of a physics course unit of the University of Torino: `https://goo.gl/pDeTw1`.

(TUM), the Karlsruhe Institute of Technology (KIT), Ruprecht Karls Universität Heidelberg and Georg-August-Universität Göttingen. The choice of these 4 universities is also consistent with the Italian–English data set, where texts from two institutes of technology and two universities were included in the data set. As previously mentioned, these websites were chosen based on the following requirements: good-quality alignable texts, large amount of texts, and ease of extraction. More to the point, in the KIT website only bilingual CUDs were available, while degree programmes are only described in German. From the TUM website both CUDs and degree programmes were published. Ruprecht Karls Universität Heidelberg and Georg-August-Universität Göttingen only publishes bilingual CUDs.[12]

KIT is a technical education and research centre, whose disciplinary domains include economics, informatics, chemistry and the biosciences. In particular, the website of the department of economics provides students with a complete archive of course catalogues in German and English.[13] TUM is an education and research institution with 14 departments belonging to the following macro-domains: life and health sciences, engineering and architecture, mathematics and natural sciences, and social sciences. In this case, a list of bilingual CUDs was available for the department of informatics.[14]

The University of Göttingen offers many degree programmes in disciplines of various kinds (social sciences, exact sciences, languages, law) for which a description is provided both in German and in English.[15] Bilingual DPDs from Heidelberg University were also available and complied with the criteria listed in Sect. 5.2.1.

Some of the other top-ranked university websites according to Webometrics were not exploited to collect texts because they did not meet the requirements set in Sect. 5.2.1. As a matter of fact, even some department websites of the KIT and the TUM did not contain any bilingual course unit description.

Ludwig-Maximilians-Universität München is ranked among the top universities in the Webometrics website. This university provides an international website which is entirely in English, but its contents are not alignable with those of the German website. A similar case is the University of Bonn, whose website – and that of its departments – is partially translated into English, but DPDs and CUDs are not. The same goes for Freie Universität Berlin, which only provides brief descriptions that are often not – or only partially – translated into English.

Similarly to what was observed for the University of Torino, also the University of Freiburg provides translated versions of its degree programmes, but since they are split into many tabs, collection would be a time-consuming task.

Universität Mainz provides a detailed description of its degree programmes in German, but only part of it is translated into English.[16] Course catalogues are available in its department websites, but their English versions are not available.[17]

5.2.4 Parallel corpora

Sentence pairs were extracted from the 4 Italian and German university websites listed in Sect. 5.2.2 and 5.2.3, automatically aligned and manually revised (see Sect. 5.2.1 for

[12] This information on the four websites was last checked on 1st February 2018.
[13] `https://www.wiwi.kit.edu/Archiv_MHB.php`.
[14] `https://bit.ly/2Mi8iSc`
[15] `https://www.uni-goettingen.de/en/3811.html`
[16] `http://www.studium.uni-mainz.de/studienfaecher-ba/`
[17] `http://www.phmi.uni-mainz.de/2944.php`

the method used). After download, each text pair was aligned using LF Aligner.[18] The aligned sentence pairs of each university were then manually reviewed in order to get rid of noisy sentences – e.g. containing XML tags or characters corrupted due to wrong encoding –, misalignments, low-quality source or target texts and so forth. For example, many target sentences were only a brief summary of their source. In other cases, typos, or linguistic issues were found in either the source or the target side. All these sentences were removed. A cleaning step using TMOP (Jalili Sabet et al., 2016) was also carried out to remove sentence pairs with corrupted characters, XML tags or wrong alignments possibly not identified during the manual process.

In order to obtain a domain adaptation set and a test set for each language combination, the corpora were split adding ca. 5% of the total number of sentences extracted from each website to the test set and the remaining ones to the domain adaptation data set. Sentences added to the test set were randomly chosen among unique sentences with a length higher than 5 tokens.

In Tables 5.1, 5.2, 5.3, 5.4 the main characteristics of the domain adaptation and test data sets for each language combination are outlined. Rows are grouped based on the text type (CUD or DPD). Besides the amount of sentence pairs, for each row the number of tokens and the vocabulary size are described. Token counts represent the number of occurrences of each linguistic unit (e.g. a word, a punctuation mark, etc.) in the data set. Vocabulary here is intended as the type count, i.e. the number of units occurring at least once in the data set. The ratio between the number of types and tokens in one or more texts is called type-token ratio (TTR) and measures the lexical variability of that text (Baker, 2010, pp. 19–21). The higher the TTR value, the higher the amount of different words. TTR is reported in the last column.

Italian–English domain adaptation set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			It	**En**	**It**	**En**	**It**	**En**
CUD	Unibo	28,406	389,189	372,175	31,728	28,381	0.08	0.08
	Polito	2,863	45,238	43,445	6,896	5,887	0.15	0.13
DPD	Polito	6,522	126,529	119,301	6,299	5,149	0.05	0.04
	Unibo	1,453	39,209	37,013	2,871	2,413	0.07	0.06
	UniRoma2	737	18,899	16,762	3,365	2,639	0.18	0.16
	Polimi	380	13,159	12,540	2,544	2,109	0.19	0.17
	Total	40,361	632,223	601,236	53,703	46,578	0.08	0.08

Table 5.1: Amount of sentence pairs and tokens, vocabulary size and TTR for each of the It–En domain adaptation corpora divided by their text type (DPD or CUD).

The domain adaptation and test sets shown in Tables 5.1 and 5.2 were obtained for It–En at the end of the data collection and cleaning process. Type-token ratio (TTR) for the whole domain adaptation data set is the same for Italian and English (0.08). In the test set, TTR is slightly higher for Italian (0.28) than for English (0.26), and in general TTR is higher than in the domain adaptation data set. These numbers – especially those for the domain adaptation data set – clearly show that sentences in these corpora are highly repetitive. Despite being richer from a morphological point of view, Italian has the same TTR as English in the domain adaptation data set. An explanation for this might be that

[18] https://sourceforge.net/p/aligner/wiki/Home/

Italian–English test set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			It	**En**	**It**	**En**	**It**	**En**
CUD	Unibo	1,697	26,952	25,808	6,389	5,791	0.24	0.22
	Polito	182	3,107	3,042	1,236	1,117	0.40	0.37
DPD	Polito	160	3,488	3,247	1,124	950	0.32	0.29
	Unibo	60	1,326	1,292	600	513	0.45	0.40
	UniRoma2	41	915	819	475	405	0.52	0.49
	Polimi	17	374	381	236	234	0.63	0.61
	Total	2,157	36,162	34,589	10,060	9,010	0.28	0.26

Table 5.2: Amount of sentence pairs and tokens, vocabulary size and TTR for each of the It–En test corpora divided by their text type (DPD or CUD).

the Italian sentences used in these texts are repetitive and often short or characterised by a simple structure. In such sentences it is therefore unlikely that morphological differences between Italian and English result in higher TTR for the former language than the latter. In general, TTR is higher for the test set, since this is composed of unique sentences.

Looking at the single rows of Table 5.1, it is interesting to notice how for one of the corpora the TTR is particularly low, i.e. DPD Polito (0.05 for italian and 0.04 for English). As a matter of fact, DPDs extracted from the Polito website were rather repetitive texts with a standard structure, where for example one paragraph described the skills to be acquired and another one listed career opportunities. This probably led to a rather limited vocabulary.

Comparing the figures in Tables 5.1 and 5.2 to those in Tables 5.3 and 5.4, the difference in the total amount of sentence pairs and tokens between It–En and De–En is clearly visible. However, no previously built parallel corpora were available for the latter language combination. Considering the time needed to align and manually review this amount of sentence pairs, it was chosen to first experiment on both language combinations in order to decide if a higher quantity of German–English bilingual texts was needed.

German–English domain adaptation set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			De	**En**	**De**	**En**	**De**	**En**
CUD	KIT	11,682	131,485	152,861	10,837	7,757	0.08	0.05
	TUM	4,457	33,638	40,221	4,413	3,368	0.13	0.08
DPD	Heidelberg	1,148	16,163	19,423	3,729	2,938	0.23	0.15
	Göttingen	819	10,607	12,613	2,440	2,011	0.23	0.16
	TUM	748	12,312	14,667	2,020	1,693	0.16	0.11
	Total	18,854	204,205	239,785	23,439	17,767	0.11	0.07

Table 5.3: Amount of sentence pairs and tokens, vocabulary size and TTR for each of the De–En domain adaptation corpora divided by their text type (DPD or CUD).

TTR for the domain adaptation set (Table 5.3) (0.11 for De and 0.07 for En) is higher than that of the test set (Table 5.4) (0.30 for De and 0.23 for En). As in the previous section (see Sect. 5.2.4) these figures can be explained with the fact that the test set is composed of unique sentences. A higher TTR and a larger vocabulary are thus expected. Comparing the two languages, German is morphologically more complex than English

German–English test set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			De	**En**	**De**	**En**	**De**	**En**
CUD	KIT	747	12,697	14,406	3,089	2,545	0.24	0.18
	TUM	270	4,945	5,593	1,637	1,424	0.33	0.25
DPD	Heidelberg	75	1,352	1,658	654	661	0.48	0.40
	Göttingen	53	947	1,174	448	424	0.47	0.36
	TUM	36	851	1,006	455	450	0.53	0.45
	Total	1,181	20,792	23,837	6,283	5,504	0.30	0.23

Table 5.4: Amount of sentence pairs and tokens, vocabulary size and TTR for each of the De–En test corpora divided by their text type (DPD or CUD).

and characterised by a large use of compounds, which decreases the number of tokens and increases the amount of types. Moreover, CUDs are often written by native speakers of the source language (Fernandez Costales, 2012), who might have a rather limited vocabulary in the target one. It is therefore not surprising that TTR is higher for German.

Average TTR for DPD corpora in the domain adaptation set (0.21 for German and 0.14 for English) is higher than average TTR CUD ones (0.10 for German and 0.06 for English). This is due to the characteristics of the two text types, which has an influence on their style and lexical variability. CUDs address enrolled students, while DPDs include features that are typical of promotional texts. Their style is thus more important. Also, according to the ECTS Users' Guide DPDs should include more information than CUDs, which might contribute to increase vocabulary and TTR (see also Chapter 2).[19]

5.2.5 Summary

This section analysed the process through which the two parallel corpora for Italian–English and for German–English were created. While for the former combination a reasonable amount of aligned texts was already available, for the latter the process began from scratch and a lower amount of sentence pairs is thus available. Evaluation results in the next sections are expected to prove if this size is enough to obtain satisfying results.

Interesting information on the texts handled here was already provided by the TTR for each of the corpora, which show that institutional academic texts are characterised by a rather low degree of variability. After this preliminary information on the corpora built for the present contribution, in the following sections MT is applied and its performance is evaluated according to different metrics.

5.3 Overall MT quality evaluation

5.3.1 Evaluation scenarios

Given the novelty of the application of MT to the translation of course catalogues and the lack of high-quality bilingual alignable texts, MMT and GT are exploited as MT systems and two realistic scenarios for one or more universities willing to start using MT are inspected. The rationale for the choice of these two systems was provided in Sect. 5.1.

[19]The ECTS Users' Guide can be found here: `https://bit.ly/2ngHclW`.

- **First scenario (GT, MMT-I)**. One or more universities want to use MT for the translation of their course catalogues for the first time. At this point, in-domain bilingual texts are not available.
- **Second scenario (MMT-II)**. A university consortium agrees to coordinate their communication strategies. They use CAT tools for translating their course catalogues and produce a reasonable amount of translations, which can be leveraged as shared domain adaptation data.

In order to address the second scenario, the domain adaptation data sets described in Tables 5.1 and 5.3 were used. The test sets in Table 5.2 and 5.4 were translated.

Since the online generic version of GT is not adaptive, it can be tested in the first evaluation scenario only. As a SOTA system, GT provides an external validation of the quality of MMT. MMT is instead evaluated in both scenarios, to analyse the impact of in-domain data on translation quality.

After this first two scenarios, a third and additional one was tested as well. In this case, the aim was to investigate the feasibility for a new university – which has no translated data yet – to exploit course catalogues already translated by other universities. Since this scenario leverages the data introduced in Sect. 5.2.4 and 5.2.4, but with different settings, corpora and results are presented in a separate section.

5.3.2 Metrics

A first evaluation exploiting two automatic metrics, i.e. the BLEU score (Papineni et al., 2002) and CHARCUT (Lardilleux and Lepage, 2017) was carried out (see Sect. 3.4 for information on these metrics). BLEU was chosen since, being widely used by the MT research community, it provides a reliable benchmark for the quality of the trained engines (despite its limitations described in Sect. 3.4). On the other hand, CHARCUT scores are provided because character-based metrics have shown to be better correlated with human judgements than BLEU and TER (see Sect. 3.4). Also, since NMT works on a subword level, character-based metrics should better take into account differences between the reference text and the hypothesis one (Lardilleux and Lepage, 2017; Way, 2018).

5.3.3 Evaluation results

Results for Italian–English (see Table 5.5) show that MT can be helpful in the first scenario already, i.e. where only generic systems are available. Performances of MMT-I and GT are similar both in terms of BLEU and CHARCUT, which confirms that the systems chosen have comparable performances. Going from the first to the second scenario, it is particularly encouraging to notice the increase of 7.71 BLEU points between MMT-I and MMT-II (-3.17 according to CHARCUT). Despite the lack of standardisation in institutional-academic texts, domain adaptation can still bring a qualitative improvement.

For German–English as well, results can be deemed positive even in the first scenario. MMT-I and GT achieves acceptable performance according to both metrics (between 31.13 and 32.47 BLEU and between 33.48 and 33.61 CHARCUT). When the domain adaptation mechanism is leveraged, quality improves dramatically, i.e. +19.06 BLEU and 8.55 CHARCUT. This might reveal a higher degree of similarity between German texts from different universities than between Italian texts, which makes the use of domain

	It–En		De–En	
	BLEU (↑)	**%CHARCUT (↓)**	**BLEU (↑)**	**%CHARCUT (↓)**
GT	36.90	29.49	32.47	33.61
MMT-I	35.45	30.30	31.13	33.48
MMT-II	43.16	27.13	50.02	24.93

Table 5.5: BLEU and %CHARCUT scores for both scenarios, i.e. GT, MMT-I (static) and MMT-II (adapted). For consistency with the other metric, CHARCUT is presented as a percentage score (instead of its original 0 to 1 score). CHARCUT measures the edit-distance between candidate and reference translations, thus the lower its score, the better the quality. Conversely, BLEU is based on matches at the n-gram level, thus the higher the better.

adaptation more effective in spite of the limited number of available in-domain sentences. Another explanation might be that the content of texts from the same university is more standardised and consistent, and for this reason being able to leverage sentences from one university both in the domain adaptation and in the test set yields a better performance. This aspect is further discussed in Sect. 5.3.4.

Results observed for German–English are especially interesting considering the lower amount of available sentence pairs with respect to the other language combination. As stated in Sect. 5.2.4, this first generic evaluation was expected to show if more in-domain sentence pairs were needed. Results described in this section demonstrated that for German–English the amount of sentences outlined in Table 5.3 are enough to obtain a performance boost from the first to the second scenario.

5.3.4 Additional scenario

As introduced in Sect. 5.3.1, NMT in the institutional academic domain was applied and evaluated in a third scenario as well. The aim in this case was to investigate the feasibility for a new university – which has no translated data yet – to exploit course catalogues already translated by other universities.

To evaluate system performance in this scenario, the same corpora as those described in Sect. 5.2.4 were used, but with a different setting. In the previous two scenarios 5% of the sentence pairs collected for each university was set aside to build the test set. In this scenario corpora of each university are not split. Instead, one of the corpora is removed from the domain adaptation data set and used as test set. For both language combinations, the text type is the CUD.

Parallel corpora TTR is 0.08 for the Italian side of the domain adaptation data set and 0.07 for the English one. The test set has a slightly higher variability degree, since TTR is 0.14 and 0.13 for Italian and English respectively. However, using sentences from one university only causes the current test set to have a lower TTR than the one used in the previous scenarios (see Sect. 5.2.4). This is further taken into account in Sect. 5.4.2.

The German–English data set is smaller than the Italian–English one. In spite of this, it was found to yield good performance in the previous scenarios (see Sect. 5.3.3). TTR for the German side of the domain adaptation data set is 0.11 and for the English side 0.07, while it is 0.18 and 0.12 for German and English respectively in the test set. As

Italian–English domain adaptation set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			It	**En**	**It**	**En**	**It**	**En**
CUD	Unibo	30,103	416,141	397,983	32,787	29,349	0.08	0.07
DPD	Polito	6,682	130,017	122,548	6,419	5,240	0.05	0.04
	Unibo	1,513	40,535	38,305	2,918	2,446	0.07	0.06
	UniRoma2	778	19,814	17,581	3,454	2,695	0.17	0.15
	Polimi	397	13,533	12,921	2,592	2,151	0.19	0.17
	Total	39,473	620,040	589,338	48,170	41,881	0.08	0.07

Table 5.6: Amount of sentence pairs and tokens, vocabulary size and TTR in each of the It–En domain adaptation corpora used the additional scenario divided by their text type (DPD or CUD).

Italian–English test set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			It	**En**	**It**	**En**	**It**	**En**
CUD	Polito	2,365	44,802	43,116	6,639	5,626	0.15	0.13

Table 5.7: Amount of sentence pairs and tokens, vocabulary size and TTR for the It–En test corpus of the additional scenario.

observed for Italian–English, the use of sentence pairs from a single university reduces the vocabulary with respect to the domain adaptation set in Table 5.3.

Interestingly enough, as noted in Table 5.3, a higher number of English tokens with respect to the German ones correspond to a larger vocabulary for German. This might be explained by the fact that these texts are often written by native speakers and by the high number of compounds in the German language (see Sect. 5.2.4).

German–English domain adaptation set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			De	**En**	**De**	**En**	**De**	**En**
CUD	KIT	12,429	144,182	167,267	11,447	8,049	0.08	0.05
DPD	Heidelberg	1,223	17,515	21,081	3,929	3,075	0.22	0.14
	Göttingen	872	11,554	13,787	2,545	2,080	0.22	0.15
	TUM	784	13,163	15,673	2,174	1,800	0.16	0.11
	Total	15,308	186,414	217,808	20,095	15,004	0.11	0.07

Table 5.8: Amount of sentence pairs and tokens, vocabulary size and TTR in each of the De–En domain adaptation corpora used in the additional scenario divided by their text type (DPD or CUD).

Evaluation Results are shown in Table 5.10. For Italian–English GT achieved a BLEU score of 37.88 (28.12 for CHARCUT) while the adapted MMT system reached 36.33 (29.18 CHARCUT). Unfortunately the adapted version of MMT – while achieving a high BLUE score of 36.33 – did not improve over its generic version (36.26) and did not reach the GT performance of 37.88. The output quality does not benefit from the domain adaptation mechanism. This can be explained with the differences between texts collected

German–English test set								
Text	**Corpus**	**Sent. pairs**	**Tokens**		**Vocabulary**		**TTR**	
			De	**En**	**De**	**En**	**De**	**En**
CUD	TUM	1,296	23,908	26,421	4,505	3,279	0.19	0.12

Table 5.9: Amount of sentence pairs and tokens, vocabulary size and TTR for the De–En test corpus of the additional scenario.

from different universities. As a matter of fact, lack of terminology harmonisation (see Sect. 2.1) and non-compliance with institutional academic communication standards (see Sect. 2.2) are two well-known issues for course catalogues. MMT performs fine-tuning at translation time retrieving from the domain adaptation data set a batch of sentences (if any) that are similar to the sentence that has to be translated. In this case, the domain adaptation mechanism is not triggered, which means that sentences from the test set (Polito) diverge from those in the domain adaptation set. Also, it is worth noting that Polito DPDs were present in the domain adaptation data set while Polito CUD texts were in the test set. The fact that no improvement was brought by domain adaptation indicates that, in order to benefit from domain adaptation in this scenario, a domain adaptation corpus for each text type would be needed.

	It–En		**De–En**	
	BLEU (↑)	**%CHARCUT (↓)**	**BLEU (↑)**	**%CHARCUT (↓)**
GT	37.88	28.12	35.76	29.31
MMT-I	36.26	29.21	35.95	29.68
MMT-II	36.33	29.18	35.69	29.67

Table 5.10: BLEU and %CHARCUT scores for each engine, i.e. GT, MMT-I (static) and MMT-II (adapted) in the additional scenario. For consistency with the other metrics, CHARCUT is presented as a percentage score (instead of its original 0 to 1 score).

For German–English as well results are not satisfactory. MMT-II (35.69 BLEU score and 29.67 CHARCUT) is outperformed by both GT (35.76 BLEU score and 29.31 CHARCUT) and by MMT-I (35.95 BLEU and 29.68 CHARCUT). Once again it is possible to conclude that the similarity between texts from different universities – or between different texts by the same university – is too low to trigger the domain adaptation mechanism. These results show that the dramatic improvement seen in Sect. 5.3.3 was due to similarities between the texts from the same university, since this scenario has proven that texts from different universities diverge too much from one another, making it impossible to leverage existing bilingual data from different universities. Besides, even texts from the same university differ if they are not of the same type. TUM CUDs composed the test set, while TUM DPDs appeared in the domain adaptation data set. Nonetheless, the domain adaptation mechanism was not beneficial in terms of quality. The same was observed for the previous language combination. This shows that, to achieve better quality using in-domain data, these must be very similar to those that have to be translated, i.e. be of the same type and belong to the same university.

However, results for both language combinations might also be affected by the relatively small size of the domain adaptation data sets. A possible further test in this additional scenario might include more data, to better understand if a higher number of

sentence pairs – although coming from universities different from the one in the test set – can increase the impact of the domain adaptation mechanism.

5.3.5 Discussion

Results described in the previous sections (5.3.3, and 5.3.4) are especially helpful to understand the feasibility of applying (N)MT to institutional academic texts. Indeed, automatic scores show that even without leveraging in-domain data, universities might benefit from the integration of MT into their translation pipeline. Such results are even more encouraging considering the improvement MMT-II showed with respect to GT and MMT-I even though the quantity of resources to exploit was low (see Tables 5.1 and 5.3). This is especially true for De–En, where leveraging 18,854 in-domain sentence pairs, a performance increase of 18.89 BLEU points and a decrease of 8.55 CHARCUT points is observed. In Sect. 5.2.4 it was maintained that the decision of collecting more bilingual texts for De–En would have been taken after the first tests. The positive results confirm that the amount of texts collected so far are sufficient given the preliminary nature of this study.

On the other hand, the third additional scenario has confirmed that texts from different universities diverge too much from each other for the domain adaptation mechanism to be useful. Further tests in this scenario with larger domain adaptation corpora for each university might confirm or not this assumption. Apart from that, results for this scenario do not add anything new to what was already observed in the previous two. As a matter of fact, since domain adaptation is not triggered, this additional scenario overlaps with the first one where a domain adaptation data set is not available. More to the point, in the first scenario the degree of lexical variability and the number of domains covered was higher, since texts to be translated were extracted from different university websites. In this additional scenario, the data set is composed of texts from one university only, which makes results less meaningful.

Considering the complexity of institutional academic texts, the fact that they contain a high number of terms – that can also be rare for domains like astrophysics or biomedicine –, and the lack of previous work in this field, a more fine-grained evaluation was deemed helpful to understand the quality that can be expected in each scenario. The next sections focus on term translation, since – given the high-density of multi-domain terminology – this can be one of the main factors affecting domain-specific text quality and, more specifically, affecting post-editing effort.

5.4 The MAGMATic data set and terminology evaluation

5.4.1 Introduction

Besides being one of the linguistic aspects (N)MT struggles with the most, terminology is arguably key for this technology and for the language industry in general. Nonetheless, in the majority of works addressing domain adaptation, the assessment step takes place measuring the engine's overall performance on a domain-specific test. Very few studies specifically focus on the engines' ability to translate domain-specific terminology, thus resorting to test sets in which terms are annotated (see Sect. 3.5.2 and Sect. 3.5.3).

Next sections focus on the procedure developed to build MAGMATic (Multi-domain Academic Gold Standard with Manual Annotation of Terminology). MAGMATic is an

Italian–English benchmark allowing MT evaluation focused on terminology and represents one of the main contributions of the present work. The data set is composed of parallel sentences extracted from CUDs and DPDs (see Table 5.2 for the composition of the data set) where terms were manually annotated in two steps. In the first one terms on the target side were identified, and in the second one each of them was assigned to its domain. The data set includes 2,055 sentence pairs and 7,517 annotated terms. Annotating a data set containing texts from more universities rather than a test set composed of texts from one institution only – as the one in Table 5.7 – allowed to cover a higher number of domains, thus making the final data set more versatile. As a matter of fact, MAGMATic covers 22 domains. The reliability of the annotation process was measured comparing annotations by two different annotators. MAGMATic was then exploited to evaluate GT, MMT-I and MMT-II (see Sect. 5.3.1).

Data annotation Two expert linguists with a background in translation studies took part in the annotation: one of them annotated the whole data set and the other annotated a portion of it so as to allow inter-annotator agreement (IAA) assessment (see details in Section 5.4.3). Two main annotation tasks were performed on the English target side of the data set, namely *(i)* the identification of the terms and *(ii)* their classification into domain categories. In order to ensure annotation quality and comparability, guidelines were created, tested in a pilot study and then given to the annotators. Annotation guidelines are available in the appendices of the present work (see Appendix A).

Term identification Both single-word (SW) terms – i.e. terms formed of one word – and multi-word (MW) terms – i.e. terms formed of two or more words – were annotated. Since instances of language for general and specific purposes often blur into each other, making the decision as to what belongs to one or the other is prone to subjectivity bias. For this reason, annotators were asked to report on their level of confidence, distinguishing between `sure` terms and `possible` terms, the latter accounting for expressions whose terminological status and specialisation were uncertain. For example, in a description of a course on electronics, *RC-circuit* was identified as a `sure` term and *charge* as a `possible` term. Where contents of a course on chemistry were outlined, *analysis* was categorised as `possible` and *pollutants formation* as `sure`. In sentences describing teaching and evaluation methods, *exam* and *lecture* were labelled as `sure` terms, while *topics* and *notions* were labelled as `possible`. This additional annotation level is particularly useful since it supports flexible evaluation designs.

Domain annotation The identified terms were assigned to one of the following categories:

- `Disciplinary`: the term belongs to a disciplinary domain – e.g. *chemical reaction, linear equation, cholinesterase*.
- `Education`: the term belongs to the educational domain – e.g. *module, course, lecturer*.
- `Education equipment`: the term refers to educational equipment that could also be used elsewhere – e.g. *overhead projector, desk*.

While the education and education equipment categories are univocal, the disciplinary category encompasses multiple domains, i.e. multiple scientific disciplines. As will be remembered in Sect. 5.4.2, to assign each term to a specific discipline, the names of the degree programmes included in the data set were leveraged: each sentence in the data set was automatically labelled with its corresponding degree programme name and all the terms annotated as disciplinary in those sentences during the annotation process inherited the sentence domain label by default. Annotators were shown this domain label during the annotation process and asked to signal cases where a discrepancy between the label assigned automatically and the actual domain of one or more terms was observed. In these cases, annotators were asked to manually assign a different label to the term(s), selecting it from the list of degree programme names.

The annotation was carried out using the MT-EQuAL annotation tool (Girardi et al., 2014). For each English sentence, the MT-EQuAL interface displays the source sentence and the disciplinary domain. Furthermore, the tool allows the annotators to perform the two annotation steps simultaneously: they mark each term and annotate it (with the sure/possible distinction and domain category) in a single go. This makes the annotation task efficient and less effortful.

5.4.2 Annotation statistics

In 101 sentences out of 2,157 (see Table 5.2) no terms were found. At the end of the process MAGMATic included 2,055 sentence pairs and a total of 7,517 term tokens, which correspond to 5,132 term types. Details regarding the number of terms annotated in the data set are provided in Table 5.11.

The disciplinary category is the largest, while the education equipment category is the smallest. Looking at the proportion between sure and possible terms for each category, it is interesting to note that possible terms are much more frequent in the education category (27.2% of the total terms) than in the disciplinary (12%) or education equipment (15.9%) categories. It can be assumed that disciplinary or education equipment terms are rarely encountered in everyday language, and are thus easier to identify as terms. On the other hand, education-related terms are also used outside of the domain, making the decision as to their status more difficult.

	Disciplinary		**Education**		**Equipment**		**Total**
	Sure	**Poss.**	**Sure**	**Poss.**	**Sure**	**Poss.**	
SWs	2,298	295	868	323	111	21	**3,916**
MWs	2,464	359	491	186	85	16	**3,601**
Total	4,762	654	1,359	509	196	37	7,517
	5,416		**1,868**		**233**		**7,517**
Vocabulary	**4,316**		**686**		**130**		**5,132**

Table 5.11: Number of terms annotated in the MAGMATic data sets. Terms in the three domain categories – Disciplinary, Education, Education-equipment (here Equip.) – are further split into the Sure and Possible (Poss.) subcategories. For either of these subcategories, the number of SWs and MWs, and the total number of terms are provided. In the bottom two rows, the total number of terms and the vocabulary (i.e. the number of types) are given for each category.

Looking at SWs and MWs, their number in the data set is approximately the same.

Domain	SWs	MWs	Total
Chemistry	345	367	712
Informatics	256	224	480
Physics	184	283	467
Biology	245	212	457
Mechanical engineering	200	210	410
Medicine	233	186	419
Electrical engineering	145	198	343
Economics	148	161	309
Mathematics	114	188	302
Environmental engineering	132	138	270
Civil engineering	109	109	218
Pharmacy	97	115	212
Statistics	90	116	206
Zootechnics	70	53	123
Aerospace engineering	65	51	116
Geosciences	62	47	109
Industrial engineering	48	59	107
Astronomy	21	61	82
Law	15	34	49
Institutions	14	11	25
TOT	**2593**	**2823**	**5416**

Table 5.12: List of the macro-domains in the data set (from the most to the least populated) and number of terms in each of them (SW, MW and total).

However the `disciplinary` category contains more MWs than SWs, whereas for the two other categories the opposite is the case. This is in line with what was stated above, i.e. `disciplinary` terms are highly domain-specific, and thus more likely to be MWs than, for example, `education` ones. The average length of MW terms is 2.44 words. Comparing the number of term occurrences with the corresponding vocabulary, terms in the `education` category show a much lower degree of variation than `disciplinary` terms. Indeed, TTR amounts to 0.80 for the `disciplinary` category, 0.37 for `education` and 0.56 for `education equipment`. This is due to the fact that the `disciplinary` category includes multiple domains, and thus a high number of different terms, while `education` and `education equipment` terms are stable and repeated across most texts. Also, the 5 most frequent terms in the data set belong to the `education` category (SWs: *student, course, students, knowledge, lectures*; MWs: *oral exam, end of the course, written test, oral examination, written exam*).

As concerns the specific domains represented in the `disciplinary` category, the specific domain labels were assigned to the terms by exploiting the names of the degree programmes of the universities from which the data set was derived (see Sect. 5.4.1). These names refer to domains with different granularity – e.g. biology and biotechnology – and thus different size. To obtain a more homogeneous set of domains, the most specific ones were merged with the generic ones where appropriate, e.g. biotechnology was grouped with biology and biomedicine with medicine. This procedure resulted in 20 macro-domain labels with a similar level of granularity. It is also worth noting that one of the reasons why it was possible to cover such a good number of different domains was

the choice to annotate a test set where texts came from different institutions. For example, sentences extracted from the two institutes of technology websites largely contributed to the increase of the engineering domains.

The complete final list of macro-domains is given in Table 5.12. As can be seen in the table, the number of terms included in the most populated domains allow for an extremely thorough terminology evaluation. Also, 9 domains out of 20 include more than 300 annotated terms. Regarding the less populated domains, they are often covered by some of the most used corpora leveraged for the creation of MT engines and only three of them contain less than 100 annotated terms.[20] As stated in Sect. 5.3.4 and Sect. 5.3.5, the choice of using a corpus composed of texts coming from different universities allowed us to cover a large number of disciplinary domains with a reasonable amount of terms each.

5.4.3 Inter-annotator agreement

In order to assess the reliability of the annotations, 220 sentences – corresponding to 10% of the data set – were annotated by a second annotator. IAA was calculated for the two types of manual annotation, namely *(i)* the identification of the terms and *(ii)* their assignment to a domain category. Agreement was computed on all the identified terms, without taking into account the `sure`/`possible` distinction.

Term identification. Two different types of agreement were calculated, to account for *complete* as well as *partial* agreement. Complete agreement refers to perfect overlap of two terms annotated by different annotators (i.e. exact match), whereas for partial agreement overlap is calculated at the level of the single words composing the term.

The agreement rates were computed using the Dice coefficient (Dice, 1945). The Dice coefficient is computed as

$$Dice = 2C/(A + B) \tag{5.1}$$

where C is the number of items annotated by both annotators, while A and B are the total number of items annotated by both annotators.[21] A Dice coefficient of 1 means that there is an overlap between the two annotations in all samples observed, while 0 means that there is no overlap at all (ibid.).

For this task, the Dice coefficient was chosen over more accurate measures that take into account chance agreement, e.g. κ, because these measure the assignment of an item to the same category by more annotators, while here the agreement is measured on the identification of an item as a term or not. According to the Dice coefficient, agreement rates are 0.69 for complete agreement and 0.79 for partial agreement, which means that, respectively, 69% and 79% of the annotated terms were marked as terms by both annotators.

Given the high number of MW terms and the strict approach used for complete agreement, results may be considered satisfactory in terms of reliability of the annotations and suitability of the annotation guidelines.

[20] Examples of corpora also covering the less populated MAGMATic domains include the JRC-Acquis corpus (`https://bit.ly/2LVqZMy`) or the Europarl one (`https://bit.ly/2MjrUp0`).

[21] Note that Dice coefficient has the same value of the F1 measure computed considering either annotator as the reference.

Domain annotation. For the subset of terms for which complete agreement between the two annotators was found (495 terms), the agreement on the assigned category label (i.e. `disciplinary, education, education equipment`) was calculated as well.

To this end, the standard *kappa coefficient* κ (in Scott's π formulation) (Artstein and Poesio, 2008) was computed. This measures the agreement between two raters, each of whom classifies N items into C mutually exclusive categories, taking into account the agreement occurring by chance.

The resulting κ value is 0.95, which – according to the standard interpretation of the κ values (Landis and Koch, 1977) – corresponds to "almost perfect" agreement.

5.4.4 Evaluation metric

This evaluation is performed in the same scenarios described in Sect. 5.3.1 and using the same experimental data set as in Sect. 5.3, i.e. with GT, MMT-I and MMT-II, where only in MMT-II a domain adaptation step is performed leveraging the data described in Table 5.1.

While the previously described evaluation was based on BLEU and CHARCUT (see Table 5.5 for results), the one outlined here focuses on terminology translation and is based on the Term Hit Rate (THR) metric (Farajian et al., 2018). THR takes in a list of annotated terms in each reference sentence and looks for their occurrence in the MT output. Then it computes the proportion of terms in the reference that are correctly translated by the MT system. An upper bound of 1 match for each reference term is applied in order not to reward over-generated terms in the MT output.

Similarly to the approach adopted for IAA (see Sect. 5.4.3), two THR types are computed: *perfect THR* – where a match is scored only if the whole reference term appears in the MT output – and *partial THR*, where the overlap between the reference terms and the MT output is calculated at the level of shared tokens. In this case, function words are removed from the MW terms in the reference, so as to avoid false positives with other function words present in the MT output. For example, if the MW *classification of living beings* is both in the gold standard and in the MT output, the *perfect THR* counts 1 match, while the *partial THR* counts 3 matches (one for *classification*, one for *living* and one for *beings*, excluding any match for the function word *of*). If the output is *classification of living creatures* no matches are found according to the *perfect THR*, while one match for *classification* and one for *living* are found according to the *partial THR*.

5.4.5 Terminology evaluation results and discussion

Perfect and partial THR scores were computed on MAGMATic for GT, MMT-I and MMT–II. Table 5.13 presents results for Perfect THR. Since MAGMATic contains both SW and MW terms, the table gives the scores for each set separately in addition to the overall score. Also, to allow a more detailed analysis of the systems' behaviour on MAGMATic terms, results are provided by domain category (`disciplinary, education, equipment`) and in terms of the `sure`/`possible` distinction.

Considering the strict parameters used to calculate perfect THR, the results shown in Table 5.13 are quite satisfactory. Regarding domain categories, all systems in all scenarios perform far better on `disciplinary` terms. As for term length, SWs are, as expected, easier to translate than MWs. The most challenging terms for all MT systems are MWs

Perfect THR									
	GT			MMT-I			MMT-II		
	Overall	SWs	MWs	Overall	SWs	MWs	Overall	SWs	MWs
All	63.72	75.43	50.98	60.97	72.98	47.90	65.33	76.07	53.65
Disc	66.80	79.75	54.91	63.94	77.52	51.47	67.74	80.03	56.50
Edu	55.62	66.33	36.78	53.32	63.48	35.45	59.28	68.01	44.61
Equip	55.78	66.96	36.76	53.31	64.10	34.96	59.11	68.40	43.32
Sure	64.95	76.26	52.76	62.43	73.91	50.06	66.58	77.05	55.30
Poss	57.25	71.20	41.35	53.25	68.23	36.18	58.75	71.05	44.74

Table 5.13: Perfect THR for GT and the 2 MMT systems. In addition to the overall scores, figures for SWs and MWs are given separately. Results are provided *(i)* for the whole data set (All), *(ii)* split according to the domain category (Disc, Edu, Equip) and *(iii)* distinguishing between `sure` and `possible` terms.

Partial THR			
	GT	MMT-I	MMT-II
All	76.68	74.91	77.23
Disc	80.40	78.83	80.64
Edu	65.33	63.13	67.49
Equip	65.63	63.30	67.13
Sure	77.74	75.94	78.07
Poss	71.27	69.68	72.96

Table 5.14: Partial THR for GT and the 2 MMT systems. Only Overall scores are reported, since matches are computed at the token level. Results are provided *(i)* for the whole data set (All), *(ii)* by domain category (Disc, Edu, Equip) and *(iii)* for `sure` and `possible` terms.

in the `education` and `equipment` categories. While these results may seem counterintuitive – given that these terms are far more common than `disciplinary` ones – it has to be noted that `education` and `equipment` categories are likely to feature inconsistencies (see Sect. 2.1), since concepts can often be referred to using different terms (e.g. *exam, examination, test* or *mark, grade*). The MT output is therefore likely to contain one correct term that does not match the one in the reference. Focusing on the first scenario, GT and MMT-I have a similar behaviour, since the differences between the two systems (ranging between 2 and 4 THR points) are constant across all the different views of the data. Two exceptions are represented by the `education` and `education equipment` MW terms, for which differences are less marked (respectively 1.33 and 1.8 THR). This seems to indicate that MMT has fewer problems than GT translating the data set terms with which all three engines struggle the most based on the THR score. At the same time, GT outperforms MMT-I by 5.17 THR in the `possible` MW category, suggesting that MMT-I probably struggles more than GT with words that might not be terms.

Comparing MMT results in the two scenarios sheds light on the specific contributions that in-domain data can bring to terminology translation. First of all, in the second sce-

nario there is an increase of the overall performance on the whole data set (+4.36 THR points). The difference with respect to the first scenario is particularly evident for MW terms (+5.75), suggesting that domain adaptation did not only influence lexical choices, but also helped the system to place terms in the correct position. As a matter of fact, partial THR results in Table 5.14 show that the performance gap between the two systems is narrower. This means that the generic and the adapted MMT systems perform similarly in the generation of the SWs composing a MW, but adapted MMT is better at generating them in the correct order. For example, in one of the segments the annotated MW *classification of living beings* was correctly generated in the second scenario, while in the first one the system produced the MW *living classification*, which is a match only in the partial THR evaluation. Finally, the biggest improvement can be found for `education` and `equipment` MW terms, which – as seen above – are the most challenging for the MT systems.

As a final observation holding for all systems in both THR evaluations, there is a clear drop in performance when progressing from the evaluation of `sure` terms to that of `possible` terms. The remarkably higher performance obtained on the most reliable terms in the data set highlights the importance of having good quality, flexible gold standards to evaluate translation of terminology.

5.5 Conclusion

This chapter analysed the whole pipeline needed to apply MT to the institutional academic domain, from the creation of the relevant data sets to the evaluation of MT engines in different settings and with different metrics. Despite the difficulties in each of these steps, results were encouraging and motivate further work in this field.

The first bottleneck for the application of MT to the translation of institutional academic texts is the lack of available high-quality bilingual resources. In addition to that, the lack of standardisation that makes text from each university different from those from other universities might exclude the process of leveraging a good amount of texts translated by other universities (this was particularly true for what was observed in Sect. 5.3.4). To conclude, these texts feature a high number of terms from different domains. Besides potentially hindering the creation of an MT system, however, such characteristics make institutional academic texts the ideal test bed for MT in low resource scenarios and, in particular, for an in-depth evaluation of the ability of NMT to handle domain-specific terminology.

Results confirmed that relying on the most appropriate SOTA technologies it is possible to overcome the issues in this field, i.e. lack of standardisation and of bilingual data, turning them into development opportunities. Focusing on two scenarios – where first in-domain resources are not available and then universities start sharing their data to build a more robust data set – improvements were testified by the results of two automatic metrics, i.e. BLEU and CHARCUT (see Table 5.5) for both Italian–English and German–English. Results are surprisingly positive for the latter language combination despite the low number of available sentence pairs (see Table 5.3).

A third scenario was considered in which data from one university only are translated (see Sect. 5.3.4). The domain adaptation data set does not contain texts that are both of the same type (i.e. CUD) and from the same university. Results showed that texts from different universities diverge too much from each other to trigger the domain adaptation mechanism. However, evaluations carried out in the first two scenarios have shown that

if one university starts to integrate MT and CAT tools into their translation pipeline, after a limited number of tasks it might be able to reuse these texts for domain adaptation, beginning to observe an increase in their output quality. Future tests might investigate the possibility of increasing the domain adaptation data set including texts from new universities. If texts from two institutions are more similar than texts from other ones, then the use of domain adaptation can be beneficial.

One of the most important contributions of the present work is the creation and use of a data set called MAGMATic to assess terminology translation for Italian–English. To build MAGMATic, the test set in Table 5.2 was manually annotated to identify target terms (both SW and MW) and assign them to their domain, i.e. `education`, `education equipment` and `disciplinary`. The latter terms were further grouped in more granular domains. The composition of the data set is described in Table 5.11 and 5.12. MAGMATic was then used to evaluate MMT-I, MMT-II and GT in the first and second scenarios. Evaluations based on THR (see Sect. 5.4.4) showed that adapting a generic NMT model to a specific domain leveraging a reasonable amount of data brings a better performance in the translation of terms (see Table 5.13). MMT-II outperforms MMT-I and GT especially on MWs and on disciplinary terms, which are the most difficult ones to handle as explained in Sect. 5.4.5.

MAGMATic is one of the few manually annotated data sets for terminology assessment (see Sect. 3.5.4), and it can contribute filling the gap in this field. This is one of the reasons why the data set was released. MAGMATic is freely downloadable from: `https://ict.fbk.eu/magmatic/` under CC BY–NC–SA 4.0.

From a more general point of view, results discussed in the present chapter have also underlined the importance of sharing data. If universities agreed to make their texts available or worked together toward developing shared (terminology) data bases, research in the translation technology field – among others – would take a leap that would mean a more efficient and effective translation for a high number of universities. This virtuous circle might bring an increased ability to communicate with foreign students, a larger web impact and a higher standardisation of texts, which in turn would make it easier to leverage translated contents from other universities to reach a better output quality.

Chapter 6

Assessing translator trainees trust towards MT

6.1 Introduction

Human trust towards MT is arguably a major factor affecting the perceived quality and the adoption or non-adoption of MT suggestions, and the choice of relying on this technology in general (see Chapter 1 and Chapter 4). Nonetheless, this factor has been largely neglected by the research community – with a limited number of exceptions (Cadwell et al., 2018; Martindale and Carpuat, 2018) – even after NMT and its fast-paced progress shook the translation industry and the research world, causing different reactions. A part of the research world has responded with enthusiastic claims about the quality achieved with this new architecture (Hassan et al., 2018; Wu et al., 2016), while other studies have tempered such enthusiasm, reporting less clear-cut improvements (Castilho et al., 2017b; Toral and Sánchez-Cartagena, 2017).

Companies and individual professionals have started to exploit MT more than in previous years. As testified by the 2018 Language Industry Survey, for the first time more than half of the participants, among which companies, independent professionals and training institutions, have stated that they use MT in their workflow.[1] On the other hand, half of the respondents still use the free and generic Google Translate, thus possibly suggesting an unwillingness to invest in MT. In the same survey repeated in 2019, only generic MT engines (Google Translate and DeepL) were again chosen among the 20 most-used tools in companies' workflow.[2]

In this uncertain scenario, translators' opinion on MT is likely to be mixed. In the 2019 Language Industry Survey, MT was identified as a negative trend by 20% and as a positive one by 30% of the respondents.[2] Lack of training in MT, low output quality resulting from adoption of general purpose engines, and a potential downward trend in translation rates may all explain the negative opinion (some) translators have of MT (Läubli and Orrego-Carmona, 2017), and their limited trust, leading to non-adoption of MT suggestions (Cadwell et al., 2018). However, there is a new generation of translators that is about to enter the market having specific knowledge on MT and post-editing. These trainees started studying translation after the neural outbreak, i.e. with an increased MT output quality, and since they have a limited professional experience (see Sect. 6.7.2),

[1] The 2018 Language Industry Survey is a survey on trends in the language industry carried out by EUATC, Elia, FIT Europe, GALA and LINDWeb: `https://bit.ly/2G0GfTR`.

[2] https://bit.ly/2ZknGlL

they are unlikely to be influenced by post-editing rates. Investigating how trust towards MT influences translator trainee behaviour towards the output, along the lines of Martindale and Carpuat (2018), is thus crucial to evaluate the likelihood that an increasing number of translators convincingly embrace MT in the near future.

The aim is to understand, through an experiment, whether translators' trust changes based on the kind of task they are working on, i.e. if they behave differently when they believe they are revising a human translation (HT) *vs.* post-editing an MT output. Here, trust is seen as strictly related to productivity: when post-editors/revisers do not trust the text they are working on, they are likely to carry out time-consuming and potentially unnecessary searches, or perform unnecessary edits.

The language combination for this experiment is Italian–English. Although translating into English as L2 is not common practice for experiments in this field, the reality of the profession is quite different. Two surveys quoted by Pokorn (2016) revealed, respectively, that for 24% of the respondents the ability of translating into an L2 is essential or important for newly employed translators[3] and that more than 50% of 780 free-lance translators working in 80 states (including Italy) translate into L2.[4] Stewart (2000, 2011) argued that Italy is among the countries where translation into L2 English is common practice.

The next section is dedicated to work on post-editing (Sect. 6.2). Even though the rest of the related work regarding MT was introduced previously, since Chapter 3 was specifically on MT architectures and how they cope with terminology, related work on post-editing is introduced here. This is meant to contribute to the clarity of the following sections, which present the method adopted for the experiments on trust.

In the remainder of the chapter, goals and variables of the experiment are outlined in Sect. 6.3, while Sect. 6.4 introduces the structure of the pilot experiment (text used, participants' background and task) and is followed by a discussion of the results (see Sect. 6.5) and of its limitations (see Sect. 6.6), which led to a new experiment, whose structure is described in Sect. 6.7. The following sections illustrate results of the final experiment (see Sect. 6.9) that are then discussed in Sect. 6.9. Participants' answers collected through a post-experiment questionnaire are provided in Sect. 6.10. To conclude, the whole experiment results and limitations are discussed in Sect. 6.11.

6.2 Related work

To the best of my knowledge, limited work has been published on the assessment of trust towards MT as measured in a PE task (see also Sect. 4.4.5). Martindale and Carpuat (2018) conducted a survey among non-professionals to understand how their trust was influenced by fluency and adequacy. The former issue is found to have a stronger negative impact on non-professional translators. More recently, Cadwell et al. (2018) interviewed two groups of institutional translators to investigate the reasons for adoption or rejection of MT suggestions. Both groups mentioned lack of trust toward MT as one of the reasons for rejecting MT segments.

Focusing on PE tasks in different languages, a number of papers have analysed how performance changes for different subjects or in different work environments, and using

[3] 2011 OPTIMALE survey, involving translation companies from 27 countries – including Italy: `https://bit.ly/2x3V0Bo`.

[4] 2014 survey by the International Association of Professional Translators and Interpreters: `https://bit.ly/2h0bjsO`.

one or more effort categories among those listed by Krings (2001): temporal, cognitive and technical. Temporal effort refers to the time needed to solve issues in an output. It includes all events occurring when working on a text, such as editing time (the time spent editing a part of the output), as well as pauses and reading. Cognitive effort is the most difficult to measure, since it focuses on the type and size of cognitive processes activated in a particular post-editing task, e.g. the number of fixations in a task or their average duration (Daems et al., 2017). Specialised tools capturing eye movements and eye fixations are often used for cognitive studies, but Think-Aloud Protocols (TAPs) – a procedure that requires translators to verbalise in real time all the processes carried out during a task – are also an option (Krings, 2001). Technical effort measures the number of operations carried out to solve the issues identified in the output. These operations can be referred to as *post-editing effort* as well (ibid.). To measure technical effort, edit-distance is widely used, i.e. the count of the edits performed to turn the raw output into the post-edited version. This count is then usually normalised on the length of the string where edits were performed. Metrics that are often used to compute the edit-distance include HTER (see Sect. 3.4), which is used in the present work (see Sect. 6.5.1 and 6.9.1). Besides edit-distance, technical effort can be determined through the number of keystrokes, mouse movements, mouse clicks or copy and paste operations carried out during a task. Keystroke logging tools are available to capture this information, e.g. Inputlog.[5]

Sánchez-Gijón et al. (2019) contrasted post-editing and translation memory (TM) editing in a blind task with 8 English–Spanish professional translators. Regarding productivity – editing time, edit-distance and amount of edited characters – only the edit-distance analysis showed significant differences between TM and MT. Quality perception was similar for the two methods. Interestingly, participants who generally perceived MT as having a positive impact on their productivity were actually a little faster when post-editing, though not significantly so.

Moorkens and O'Brien (2015) used edit distance and speed to compare the productivity of professionals and students in a PE (En–De) task, whose aim was to evaluate the suitability of the latter for translation user studies. Daems et al. (2017) examined how 10 Master's students and 13 professional translators coped with translation from scratch and PE of newspaper articles (En–Nl), measuring translation speed and cognitive load. Moorkens and O'Brien (2015) found that students have a less negative attitude towards technology, but their productivity cannot be compared to that of professionals – speed for students was less than half that of professionals –; by contrast, according to Daems et al. (2017) the performance of the two groups was not as different as could be expected – differences between students and professionals in terms of processing speed were not statistically significant –, and indeed students were more at ease with PE than professionals.

Yamada (2019) compared perceived cognitive effort, amount of editing and final quality between two PE tasks carried out by students, one using an NMT output and one a PBMT output (En–Ja). While the cognitive effort was similar for the NMT and PBMT tasks, NMT output required less editing effort and led to a better final quality.

Rossetti and Gaspari (2017) measured perceived and real effort of six MA students when translating with TMs and in a PE scenario, triangulating time measurements, think-aloud protocols (TAPs) and retrospective interviews. Results show that only suggestions coming from the TM had a positive impact on perceived task complexity and temporal effort.

Despite the amount of work on post-editing effort, results for productivity compar-

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

isons between post-editing of an MT output and revision of a HT, as well as outcomes on the differences between perceived and real effort, are often inconclusive. Besides, to the best of my knowledge trust has not been investigated in such tasks. Furthermore, the observed language combination (It–En) is relatively under-represented in PE experiments, and the text domain (university module descriptions) is a novel one in this scenario.

6.3 Goals and variables

As introduced above (see Sect. 6.1), the main goal of the experiment presented here is to understand whether translators' trust changes based on the kind of text they are working on, i.e. if they behave differently when revising a human translation (HT) *vs.* when post-editing an MT output.

Different behaviours are taken into account in two ways, based on two of the three effort categories listed by Krings (2001), i.e. technical and temporal effort (see Sect. 6.2).

- Measuring the number of edits between the original text and the post-edited/revised version using HTER (see Sect. 3.4).
- Measuring the time spent on the task computing the words per second rate.

The third category would have been the cognitive one. However, work focusing on cognitive aspects of translation and post-editing used CAT tools such as Translog (Alves and Vale, 2009; Carl, 2012; Toledo Báez et al., 2017; Vieira, 2016), CASMACAT[6] (Daems et al., 2017) or PEARL (Moorkens et al., 2015) to support eye-fixations. These tools allow for precise measurements of cognitive effort indicators, but on the other hand they come with a less user-friendly interface (except for CASMACAT). Since the focus is on trust, participants should not be negatively influenced by the work environment. Also, a termbase is used in the final experiment (see Sect. 6.7.1), and the tools mentioned above do not include the possibility of adding a termbase to the work environment.

Post-editors/revisers' temporal and technical effort are analysed with respect to the following variables: *(a)* presumed translation method (students are told that the text is an MT output *vs.* a HT); *(b)* translation correctness (the target sentence is correct and needs to be confirmed *vs.* it is incorrect and needs to be edited).

A first pilot experiment (described in Sect. 6.4 and 6.5) took place in May 2018, a second experiment was carried out in March 2019 (see Sect. 6.7, 6.9 and 6.11) and is also described in Scansani et al. (2019a).

6.4 Pilot experiment structure

6.4.1 Participants

18 first year students from the Master's in Specialised Translation of the University of Bologna were divided in two groups of 9 students each. Before starting the experiment, students filled in a questionnaire (see Appendix C) where they were asked one question on their professional experience with revision/HT, one on their professional experience with MT/PE and one question on their opinion on the usefulness of MT for translators.

[6] http://www.casmacat.eu/

Question	Answers	Participant %
Professional experience with revision/HT	None	33.3%
	Little	50%
	Much	16.7%
Professional experience with MT/PE	None	55.6%
	Little	22.2%
	Much	0%
MT usefulness for translators	Not useful	0%
	Useful	55.6%
	Very useful	44.4%

Table 6.1: Results of the questionnaire on participants' professional experience and opinion on usefulness of MT.

Questionnaire results are shown in Table 6.1. Regarding participants' professional experience, as expected only a minority of them had carried out revision/post-editing tasks outside the classroom (33.3% for revision/HT and 55.6% for MT/PE). 50% of the participants has little experience in revision/HT, and 22.2% stated the same for MT/PE. Even though no participant declared to have a good amount of experience in MT/PE, results for the opinion on the usefulness of MT for translators are positive: 55.6% think MT is useful, while 44.4% deem it very useful. These results are in line with Daems et al. (2017), reflecting a positive attitude of translator trainees towards technology.

Besides the advantages listed in Sect. 6.1, working with students belonging to the same cohort allows us to control for *(i)* their PE/translation experience; *(ii)* their knowledge of the text type and disciplinary domains of the texts; *(iii)* their knowledge of English.

Regarding *(i)*, all students attended hands-on modules on CAT tools and on MT and PE as part of their syllabus. In the module on CAT tools they learn how to use Trados Studio, which is the tool used for this experiment, both for translating and for project management. Also, results of the questionnaire listed in Table 6.1 confirm that their degree of expertise is similar.

Regarding *(ii)*, all subjects are likely to be familiar with the text type, since course unit descriptions address university students, and are unlikely to be acquainted with the domain (information technology, IT), since their academic background is in languages and linguistics. Concerning *(iii)*, all students are tested upon enrollment in the Master's, a minimum of C1 CEFR being required for admission.[7]

6.4.2 Text

The same text was used for both group A and group B. It was composed of a course unit description – for a course on IT – written in Italian. The English version was produced with the same system used for the tests in Chapter 5, i.e. the commercial version of ModernMT (MMT).[8] As introduced in Sect. 5.1, MMT is a state-of-the-art off-the-shelf NMT system, which ensures the high-quality of the target text used for the experiment (see Sect. 5.3.1).

The final version of the text was the result of a two-step procedure. First, to make

[7] `https://bit.ly/2pVyffz`
[8] `https://www.modernmt.eu/`

sure the text could be believed to be a HT, possible mistakes typical of MT systems were discarded, e.g. translation of proper nouns or issues due to a semantically wrong interpretation of the source text incompatible with human performance. To establish which sentences were (in)correct, three evaluators were asked to assign each sentence to one of the following categories:

- Wrong: the meaning of the source sentence is not conveyed in the target version.
- Incorrect: the meaning of the source sentence is conveyed in the target version, but editing is needed to achieve a high-quality version of the text.
- Correct: no editing needed.

The final decision as to the correctness of each sentence was made by majority voting. The sentences marked by the majority of the annotators as 'wrong' were discarded. Some of the other sentences were edited in order to have the same number of sentences for each of the correct and incorrect conditions.

At the end of the selection and evaluation procedure, the text had a length of 350 words ca. divided into 21 segments. 13 additional segments are provided locked – and without a translation – to offer context and make sure participants have a clear view of the whole text. Also, locked segments are added to signal the beginning/end of the HT or MT text portion (see Fig. 6.1). Four additional segments (80 words) were used in a warm-up task. Both the warm-up and the experiment texts are (part of) course unit descriptions belonging to the IT disciplinary domain.

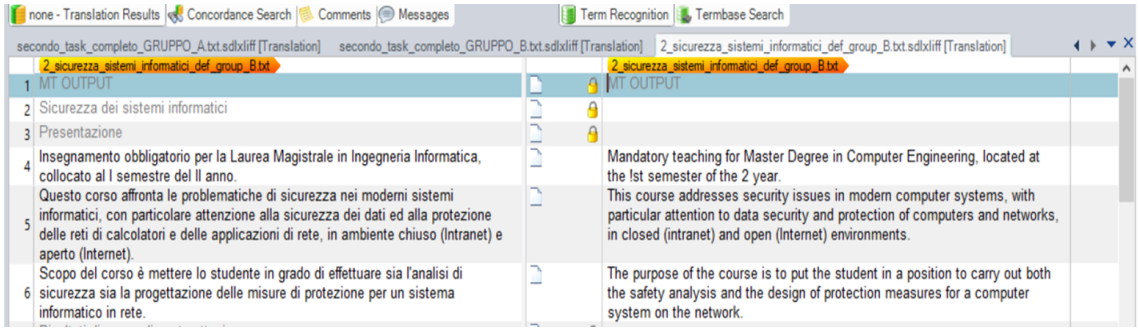


Figure 6.1: First segments of the text provided for the trust assessment task to group B. The first segment (locked) signals the beginning of the MT output.

6.4.3 Task

A week before the experimental session, students were given basic information about the experiment, i.e. that the aim was to compare PE and revision, that data would be collected anonymously and that taking part in the experiment was not compulsory.

Before the task, instructions were provided to students (see Appendix B). They were asked not to over-edit the text – using as much of the raw text as possible–, to work as they were used to, e.g. exploiting their usual resources, and to provide a high-quality publishable final version of the text. After reading the instructions, students started working autonomously. Researchers were present in the lab throughout.

The tool used for the experiment was Trados Studio. The target text was already included in the Trados package. Adding the Qualitivity plugin to the Trados environment it was possible to keep track of the time spent and the edits performed on a segment level.[9]

6.4.4 Data collection and analysis

Productivity was measured in terms of HTER (Snover et al., 2006) between the original text and the participants' edited version (see Sect. 5.4.4) and in terms of words per second (WPS). The latter is obtained dividing segment-level time measurements by the number of target words.

Two separate linear mixed models were built, one for each dependent variable, i.e. HTER and WPS. In both cases, the independent variables (and fixed effects) were categorical, i.e. translation method (MT/HT), and translation correctness (correct/incorrect). An interaction of the two was included in the model, with participant and segment as random effects.

Random effects were tested for significance using the likelihood ratio test. Following Gries (2015), a model including all fixed and random effects was built and compared using analysis of variance (ANOVA) against different null models, each excluding one of the random effects. If the difference between the two models was significant ($p < 0.05$), the random effect was kept in the model.

6.5 Pilot experiment results

6.5.1 HTER results

Before applying the analysis procedure explained in 6.4.4, outliers – i.e. observations whose HTER value was higher than the sum of the mean and 2 standard devations (SDs) were discarded. This threshold follows Ferraresi (2016), where 3 SD plus the mean was the suggested threshold for time observations. However, in that case the analysed variable was a continuous time measurement, thus more subject to fluctuations, while HTER is a percentage normalised on the number of words in the target text, thus less prone to variations.

After removing 22 outliers, the data set contained a comparable number of observations for each condition (see Table 6.5.1).

Transl. method	Transl. correctness	Observations
HT	Incorrect	86
HT	Correct	84
MT	Incorrect	85
MT	Correct	83
Total		**338**

Table 6.2: Number of observations for the HTER analysis, categorised according to each translation correctness and method conditions, after having removed the outliers.

[9]The Qualitivity plugin can be found here: `https://bit.ly/2KVviHi`.

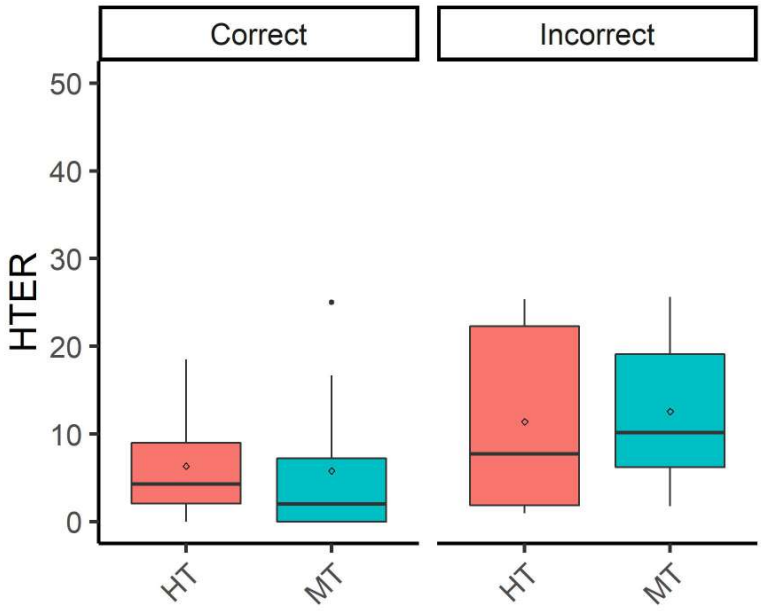


Figure 6.2: HTER values for segments split by translation method and correctness of the translation.

The linear mixed models were built with HTER as the main variable, translation method and translation correctness as independent variables, and participant and segment as random effects (see Sect. 6.4.4). Results are summarised in Table 6.3 and 6.4.

HTER		
	Effect	***p* value**
Random	Participant	<0.001***
	Segment	<0.001***
Fixed	Correctness	0.09
	Method	0.78
	Interaction	0.74

Table 6.3: Significance of random and fixed effects on the two dependent variables: HTER.

Variables	**HTER**
HT incorrect	12.19
MT incorrect	12.10
HT correct	5.49
MT correct	4.57

Table 6.4: Estimates of the two linear mixed models for HTER. HTER goes up when more edits are performed.

Both *participant* and *segment* have a significant effect on the main variable while neither of the two fixed effects, nor their interaction, have a statistically significant effect. In other words, the number of edits does not change significantly between correct and incorrect sentences and between MT and HT sentences. Also, no significant change in HTER scores is observed in HT and MT across translation correctness conditions. As a matter of fact, looking at Table 6.4, in both MT and HT conditions, HTER is similar for correct and incorrect sentences. The most marked difference is between correct and incorrect sentences.

The fact that the difference between correct and incorrect sentences is not significant may cast doubt on the appropriateness of the experiment structure. Figure 6.2 shows that for incorrect sentences there is a dramatically high variability between the observations.

The HT box covers a 20% span and its lowest value reaches almost 0, which means that some incorrect sentences were only slightly edited. Results for the correct condition, where the two boxes cover a smaller portion of the HTER values, are more similar to those expected.

Analysing HTER data on a sentence level, large differences between the first and the second half of the sentences emerged. Course unit descriptions are composed of two distinctive parts. The first one can be more complex especially from a terminology point of view, since disciplinary contents of the course unit are outlined. The second part typically describes teaching and assessment methods, topics students are more familiar with. As a matter of fact, mean HTER is 10.96 for the first half of the sentences – which includes all disciplinary related sentences – and 23.71 for the second half. Unless the output quality was substantially better in the first half of the text, which would be surprising given its complexity and the number of domain-specific terms, this seems to suggest that differences in the two parts of the text influenced participants' behaviour. If this were true, then results would be biased, especially given that each half of the text was assigned to one of the two conditions MT and HT. Sect. 6.5.3 will further investigate this issue.

6.5.2 WPS results

As in Sect. 6.5.1, before analysing the data set, possible outliers – observations higher than the mean plus 2 SDs – were discarded. The method used in this paragraph is the same as the one described in 6.4.4.

As in the HTER analysis (see Sect. 6.5.1), 22 outliers were found and removed from the data set. After removal, the number of observations across the conditions was uneven. For this reason, all observations related to two 'Incorrect' segments were removed from the data set, resulting in the distribution in Table 6.5. It has to be noted, however, that this data set included more observations than the one used for HTER, since when participants went back to one segment they had already edited, this was considered as a new observation, whereas for HTER each segment has one observation per participant only.

Transl. method	**Transl. correctness**	**Observations**
HT	Incorrect	73
HT	Correct	94
MT	Incorrect	73
MT	Correct	100
Total		**340**

Table 6.5: Number of observations for the WPS analysis, divided by each translation correctness and method conditions, after removing the outliers.

The same models as in Sect. 6.5.1 were built, this time with WPS as main variable. Results are displayed in Table 6.6 and 6.7. Differently from what happened for HTER, the random effect *segment* does not have a significant impact on the main variable, while *participant* does. The latter is thus the only random effect to be included in the model whose results are listed in Table 6.7. Regarding the fixed effects, only the interaction between translation method and translation correctness has a p value < 0.05. This means that differences between MT and HT or between correct and incorrect sentences are not statistically significant, while differences between, e.g., MT correct and HT correct are

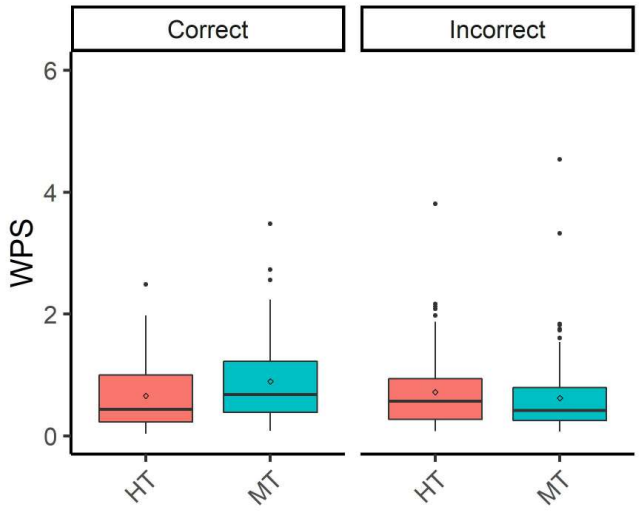


Figure 6.3: WPS values for segments split by translation method and correctness of the translation.

significant. The absence of a significant difference between correct and incorrect sentences for WPS is probably due to the fact that time measurements also include the cognitive effort, which might be the same for both incorrect and correct sentences or even higher for the former ones.

WPS		
	Effect	***p* value**
Random	Participant	<0.01**
	Segment	0.16
Fixed	Correctness	0.54
	Method	0.71
	Interaction	<0.01**

Table 6.6: Significance of random and fixed effects on the two dependent variables: HTER.

Variables	**WPS**
HT incorrect	0.82
MT incorrect	0.63
HT correct	0.48
MT correct	0.92

Table 6.7: Estimates of the two linear mixed models for WPS. When WPS increases, participants' productivity is higher.

Results in Table 6.7 are somewhat contradictory. When sentences are incorrect, participants are more productive when revising than post-editing. Conversely, for correct sentences productivity increases when the translation method is MT. Surprisingly, for HT students are more productive if sentences are incorrect, while for MT – as one would expect – their productivity is higher for correct sentences. As a matter of fact, Sect. 6.4.4 showed that in MT correct sentences a slightly lower number of edits were performed than on HT correct ones. It has to be noted, however, that since WPS is also influenced by the cognitive effort, it would be wrong to assume that a lower HTER would necessarily correspond to a higher WPS. The lower productivity on HT correct sentences might be due to the fact that HT correct sentences required a higher cognitive effort of students, who checked sentences thoroughly to make sure not to leave issues behind.

Although outliers were removed, some particularly high values are still in the data set (Fig. 6.3), with more than 3.5 WPS even for incorrect sentences. These high productivity

rates for incorrect sentences probably mean that the structure prevented participants from behaving as would have been expected in a post-editing/revision task. Also, if these high rates were not among the outliers removed from the data set, it means that there were cases in which WPS was even higher.

Unexpected behaviours for incorrect sentences, the lack of significant differences between correct and incorrect ones, and above all the differences in terms of mean HTER between the two portions of the text signal unexpected participants' behaviours, perhaps because the structure of the task influenced their activity. To gain a better insight into that, the changes made by participants were manually inspected. This is described in the next section.

6.5.3 Manual analysis

Behaviours different from those expected were observed especially in the HTER analysis, where mean HTER values diverged substantially in the two halves of the text. To investigate the reasons for such unexpected behaviours, mean HTER for each sentence in the HT and MT conditions was computed. Then sentences with HTER scores higher than the mean plus 1.5 SD were extracted from the data set and manually analysed. It has to be noted that these observations are not outliers, since outliers were removed before the analyses (see Sect. 6.5.1).

In this section, sentences are divided into those belonging to the first half of the text and those belonging to the second one. Since course unit descriptions are composed of sections (in the first half) that are more related to the discipline taught – thus containing more domain-specific terms – and other sections (in the second half) that are focused on the didactic part, this division allows to understand if different parts of the text influenced participants' behaviour. Besides, in this way it is possible to take into account differences between the behaviour of participants when post-editing and when revising, since the two conditions MT and HT were assigned to the different halves of the text.

First half 12 observations (6 HT, 6 MT). This subset is composed of 10 sentences whose raw version was correct and 2 sentences whose raw version was incorrect. All sentences were related to the disciplinary domain. Regarding the 10 correct sentences, after the edits results were still correct, but changes were not needed. The highest HTER scores (between 30 and 65) are on domain-specific sentences in which participants often changed both one term and the structure of nominal groups through substitutions and shifts (e.g. "Weaknesses of network and processing systems" changed into "Networks and processing systems' weaknesses"). Since most sentences in this subset were correct, the majority of participants did not edit them, thus even low HTER values were higher than the mean plus 1.5 SDs, e.g. "Ability to analyze the risks of a network application" changed into "Ability to analyze the risks of network applications". HTER is also influenced by the sentence length. Some of the outliers were short sentences where a few edit steps increased HTER ("Knowledge of the main types of attacks on computer systems" changed into "Knowledge of the main types of attacks against computer systems"). Regarding the two incorrect sentences, they were both produced by the same participant in the MT condition. In one case the sentence became correct after post-editing (HTER 21.4) by modifying a domain specific term, even if also unnecessary changes (replacement of "and" with commas) were introduced. The other sentence was modified through word shifts that did not correct mistakes in the sentence. As can be seen in all these examples, except

for one of the two incorrect sentences, in the other cases changes were mostly performed on non domain-specific words.

Second half 20 observations (9 MT, 11 HT). In this subset, most of the sentences (16) were incorrect before the editing steps and became correct afterwards. Since issues in these sentences were related to linguistic aspects – e.g. collocations, unclear formulations – different solutions were possible to correct the sentence. Some of them were costly in terms of number of edits, causing the HTER value to be higher than the mean plus 1.5 SD. One example is "When conducting the test, you can consult texts or notes" changed into "When carrying out the exam, refer to your books or notes". In other cases, besides introducing the necessary changes, participants changed other parts of the sentence. One example is "There are 6 to 7 different experimental exercises, with the possibility of repetition" changed into "6 to 7 different experimental exercises will be carried out and possibly repeated". Only one sentence in this set contained domain-specific terms, e.g "Information Protection Techniques: steganography, encryption" edited into "Information Protection Techniques: steganography, cryptography". Sentences were not modified by most of the participants, so even light editing resulted in a HTER score higher than the mean + 1.5 SD. In correct sentences, all of which were not related to the disciplinary domain, the HTER score ranged from 70 to 123. Some of them were radically modified through rephrasing. One example is the sentence: "The website will also provide additional material prepared by the teachers" changed into "On the website additional material collected by the teachers will be available".

Summing up results of this manual analysis, where sentences or terms are strictly related to the disciplinary domain as in the first half of the text, minor changes are introduced. When the content of a sentence is not related to the discipline, which is usually the case with the second half of a CUD, a higher number of edits or edits more costly in terms of HTER are performed, but not all of them are necessary. Confirming what was first observed in the HTER analysis (see Sect. 6.5.1), the text structure has an impact on this experiment's results, since HT and MT conditions were distributed across the two halves of the text. This will be further discussed in Sect. 6.6.

6.6 Pilot experiment – Conclusions and limitations

Results in Sect. 6.5.1 and 6.5.2 have shown that the collected observations were somewhat influenced by the experiment structure. The manual analysis described in Sect. 6.5.3 shed light on some of the reasons behind this.

The content and the degree of complexity of the two halves of the text differ from each other, which had an impact on participants edits. As a matter of fact, the highest HTER scores in the first half are about 65, while in the second half they reach 123. Also, mean HTER for the first half is 10.96, while in the second half it is 23.71 (see Sect. 6.5.1). Assigning each condition to a different half of the text as done in this pilot allows participants to follow the text structure and is useful to collect observations in both HT and MT conditions for each participant, but it has an impact on their behaviour. The first half is usually the one containing more domain-specific terminology. In this text, some sentences were particularly dense with terminology, e.g. "Information Protection Techniques: steganography, encryption, digest, X.509 certificates, certification authorities (CA), and public key infrastructures (PKI)". Sentences in the second half are easily

understandable for students. This was probably the cause for higher HTER scores in the second half of the text (see Sect. 6.5.3). Students encountered difficulties when revising or post-editing complex segments, thus they either did not notice they were wrong (or noticed it but found it difficult to correct them), or they tried to correct them, but only introducing minor changes that did not (completely) solve the issues. On the other hand, they tended to (over-)edit generic simple sentences probably because they were the ones they felt more at ease with. The reason for the over-editing might be a feeling of frustration induced by the complexity of the first sentences, where participants were not able to modify the text. Another possible explanation might be that when a sentence is more generic, students are more concerned about its quality and fluency. Solutions have to be found in order to keep the variable complexity of the text under control, which would otherwise influence observations for MT and HT and, consequently, the experiment results.

Another limitation, from a more practical point of view, was the use of Trados Studio, which proved not to be the ideal solution. First, a Trados package had to be sent to each participant, who then opened it. This extended the duration of the experiment and added an extra task to the one participants were asked to carry out. Moreover, Trados Studio does not provide built-in solutions to track productivity, and the integration of Qualitivity caused windows to automatically appear in the Trados Studio environment. Since students were not used to this, it is possible that their behaviour and their trust towards the work environment were influenced. For the final experiment, a different tool integrating productivity measurement tools and allowing for a simpler project assignment process was used.

6.7 Final experiment structure

6.7.1 Differences with the pilot experiment

To overcome the limitations described in Sect 6.5.3 and 6.6, two main solutions were adopted. MateCat[10] was used instead of Trados, and each participant was assigned to only one of the MT/HT conditions.

The structure of this final experiment prevents the collection of both MT and HT observations for each participant. However, the use of mixed models and the availability of a higher number of subjects than the 18 taking part in the pilot experiment should compensate for the absence of both MT and HT observations for each participant. Moreover, not dividing the text into HT and MT is useful to control for possible fatigue or learning effects that would degrade/enhance participants performance in the second half of the text.

Even assigning each participant to one variable only and providing them with one whole text would not solve the issue of different complexity for different sentences. Starting from the assumption that the complexity degree of a sentence and the number of necessary searches might be strongly correlated with the presence of domain-specific terms, providing participants with a validated glossary can be beneficial for students, who might also tend to perceive the task as less difficult. Making the task easier and decreasing the number of time-consuming terminology searches also makes it possible to add more sentences to the data set, thus collecting a higher number of observations, with a view to enhancing the data set reliability. At the same time, some of the sentences in the first

[10] `https://www.matecat.com/`

half of the pilot text were particularly dense with terms, which was probably one of the reasons for the low HTER in the first part (see Set. 6.6). For all these reasons, new texts were selected for the final experiment, with a view to ensuring a more careful control of complexity.

The use of an online tool such as MateCat is ideal to streamline the whole process. MateCat translation projects can be shared with a url. In this way, participants only have to open the link and start working on the text. By default, MateCat keeps track of the time spent on a sentence and includes it in a productivity report together with they key-logging for each sentence. Also, participants received training on MateCat one week before the experiment. Receiving training on a CAT tool they did not know before, might be perceived as a reward for taking part in the experiment, thus making them feel more motivated.

After having outlined the new solutions introduced for this final experiment, the next sections detail the experimental setup as previously done for the pilot experiment.

6.7.2 Participants

47 first year students from the Master's in Specialised Translation took part in the experiment in March 2019. None of them had taken part in the pilot experiment, since they belonged to a different cohort with respect to the pilot experiment participants. They were randomly assigned to the two tasks:

- 23 participants worked on the PE task
- 24 participants worked on the revision task

Native languages of the participants working on MT were Italian (69.6%), English (4.3%) and other (26.1%).[11] The native language of participants working on the purported revision of a HT was Italian (79.2%), English (8.3%) and other (15.5%).[11]

All students belonged to the same cohort. This allowed us to control for *(i)* their PE/translation experience; *(ii)* their knowledge of the text type and disciplinary domains of the texts; *(iii)* their knowledge of English.

Regarding *(i)*, students had attended hands-on modules on CAT tools and on MT and PE as part of their syllabus. One week before the experiment, they received training on the use of MateCat, the tool used for the task (see Sect. 6.7.4). Also, in a pre-experiment questionnaire (see Appendix C), they were asked how much experience they had with the revision of a HT or PEMT in a professional setting. Possible answers were: *None*, *Little*, i.e. from 1 to 5 professional tasks, or *Much*, i.e. more than 5 professional tasks. Results are reported in Table 6.8 and show that the degree of expertise is similar in both groups, since the vast majority of the participants had no or little professional experience. Also, as could be expected, expertise is slightly higher for HT than for MT.

Regarding *(ii)* and *(iii)*, motivations are similar to those listed in Sect. 6.4.1. Concerning *(ii)*, subjects are likely to be familiar with the text type and not with the domains (pharmacy and chemistry). With respect to *(iii)*, all students are tested upon enrollment in the Master's, a minimum of C1 CEFR being required for admission.[12]

To collect data on participants' opinion regarding MT, in the pre-experiment questionnaire they were asked how useful they thought MT is for translators, i.e. not useful,

[11] Languages referred to as "other" were: Spanish, French, Russian or Romanian.

[12] `https://bit.ly/2pVyffz`

Question	Answers	MT part.	HT part.
Professional experience with MT/HT	None	91.3%	95.8%
	Little	8.7%	0%
	Much	0%	4.2%
MT usefulness for translators	Not useful	0%	0%
	Useful	82.6%	70.83%
	Very useful	17.4%	29.17%

Table 6.8: Results of the questionnaire on participants' professional experience with MT/HT and opinion on usefulness of MT, split by type of task (HT or PE).

useful, very useful. Similarly to results previously seen in Table 6.1, those in Table 6.8 suggest that all participants have a positive opinion on MT, confirming the results obtained by Daems et al. (2017) and Moorkens and O'Brien (2015). According to 82.6% of the participants chosen for the MT task, this technology is useful, while 17.4% stated it is very useful. 70.83% of the participants carrying out the revision task stated MT is useful, while 29.17% stated that it is very useful.

6.7.3 Text

The same text was used for both the PE task and the revision one. It was composed of two course unit descriptions – for a course on chemistry and one on pharmacy – written in Italian. The English version was produced with MMT (see Sect. 5.3.1).

The final version of the text was the result of the same two-step procedure described in 6.4.2. None of the sentences was labelled as *wrong*. A small amount of edits were performed in order to have half *correct* sentences and half *incorrect* ones in the data set (see Sect. 6.3). At the end of this procedure, the text consisted of 60 sentence pairs, corresponding approximately to 670 source words in total.

6.7.4 Task

The tool used for this experiment is MateCat, an online CAT tool, for which students received training one week before the experiment. Reasons for this choice are provided in Sect. 6.7.1. A project containing the target text and a small termbase with 47 term pairs was created beforehand and shared with participants through a url. Students were given basic information about the experiment. They were told that the final aim was to compare PE and revision, that data would be collected anonymously and that taking part in the experiment was not compulsory.

Instructions were the same as those in Sect. 6.4.3 (see Appendix B), i.e. participants were invited to work as they normally would and to deliver a target text of publishable quality trying to use the provided target text as much as possible. After reading the instructions, participants started working autonomously. Researchers were present in the lab throughout.

6.7.5 Data collection and analysis

Time spent on a segment is measured by MateCat by default. HTER was computed after the experiment between the raw output and each edited version. The metrics used are

HTER and WPS. Metrics behaviour and significance were analysed using linear mixed models as explained in Sect. 6.4.4.

6.8 Pre-analysis sanity check

Flaws identified in the pilot experiment structure described above (see Sect. 6.6) prevented students from behaving as expected. For example, differences between correct and incorrect sentences in terms of HTER were not significant (see Sect. 6.5.1), and issues in complex incorrect sentences were often not solved. In this final experiment, before starting the result analyses, a member of the academic staff – specialised in institutional academic communication and with a background in translation – was asked to evaluate the set of sentences produced by each participant. The aim of this evaluation was to understand if students behaved as expected, thus following the instructions and producing publishable sentences for the website of a university (see Sect. 6.7.4). The evaluator was asked to label each revised/post-edited sentence with one of these three categories: *publishable*, *improvable* or *not publishable*, where the difference between *publishable* and *improvable* is that sentences assigned to the latter category are of near-publishable quality, i.e. an improvement would be welcome, for instance concerning terminological inconsistencies or stylistically questionable prepositional use.

	correctness before	**edited or not**	**publishability**	**behaviour**
1	correct	edited	publishable	ok
2	correct	not edited	publishable	ok
3	correct	edited	improvable	not ok
4	correct	edited	not publishable	not ok
5	incorrect	edited	publishable	ok
6	incorrect	not edited	not publishable	not ok
7	incorrect	edited	improvable	ok

Table 6.9: All variables considered in this sanity check are listed in this table, one for each column. All variable combinations that were found in the data set are reported here, one for each row.

Sentences that participants chose not to edit were not evaluated. They were labelled as publishable if the original version was publishable, and as not publishable if their original version was incorrect. In order to have a complete overview on what was required to participants, the evaluator had access to the instructions students read before the experiment (see Appendix B).

At the end of this evaluation, labels assigned to each sentence were intersected with the following information: *correctness before*, i.e. if the raw output was correct or incorrect (see Sect. 6.4.2 and 6.7.3) before the editing steps, *edited or not*, i.e. if students decided to edit the sentence or leave it as it was. For each of the possible combinations of these three variables (correctness before, edited or not, and publishability), a judgement on the participants' behaviour on a sentence was made, i.e. if the behaviour was *ok* or *not ok*. For example, if a wrong sentence was edited and the final result was deemed *publishable* by the evaluator, then behaviour was labelled as *ok*. Possible combinations are listed in Table 6.9.

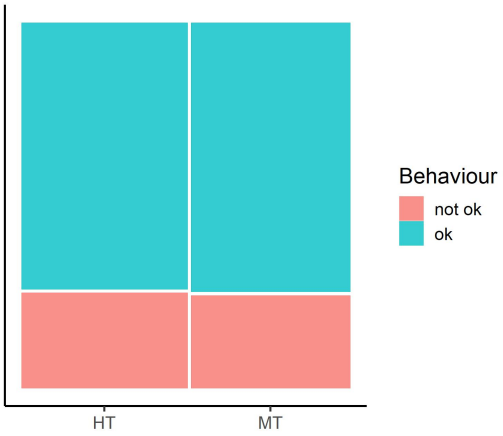


Figure 6.4: Mosaic plot showing ok and not ok behaviours for the MT and the HT condition.

First, the amount of *ok* and *not ok* behaviours was compared across MT and HT. Although using binary judgements to categorise behaviours is an approximation – e.g. when a correct sentence is edited it would be useful to know if the quality improved or not in order to decide if the behaviour was correct –, it still helps to have a general overview on participants' decision-making process, to understand if they behaved in unexpected ways in one of the two conditions. As showed in Fig. 6.4, revisers and post-editors behaved almost in the same way, which should indicate that all the possible variables in this experiment were kept more under control than in the pilot experiment, where unexpected behaviours were observed.

A portion of the behaviours was labelled as *not ok*. *Not ok* sentences include sentences that were correct before the task and became *improvable* or *not publishable* after the editing process (see row 3 and 4 in Table 6.9). From the point of view of the decision process, this is probably one of the less expected behaviours, but in the case of *improvable* sentences it is somewhat counterbalanced by the fact that the quality decrease was limited, i.e. *improvable* means that the source message is still conveyed by the target text. Regarding correct sentences that became *not publishable* after editing, these are by far the smallest subset of sentences in the data set (only 25), and one of the reasons for the non-publishability was the presence of typos in the final version of the sentence. Also, it has to be noted that correct sentences that were edited might be the proof of a lack of trust towards one of the two MT or HT conditions. The other *not ok* sentences are those incorrect that were not edited (see row 6 in Table 6.9). In other words, in these sentences participants were not able to solve (or indeed identify) issues. Since to measure trust the focus is on the process rather than the final product, the only completely negative condition is the one in row 6 (Table 6.9), where incorrect sentences were not even edited.

Summing up, the vast majority of the sentences was labelled as *ok*, and the presence of *not ok* or *not publishable* sentences, some of which do not represent an ultimately unreliable behaviour, in no way hinders the reliability of the data set. It has to be noted that a dramatically high number of *not publishable* or *not ok* sentences would have been a warning sign for the reliability of the data set, whereas a limited number of this kind of sentences might be the sign of an excess or a lack of trust towards MT or HT, which is exactly one of the phenomena this experiment aims at investigating. Besides, as it is often the case in an experimental scenario, participants might have limited experience

in the domain(s) chosen – in this case pharmacy and chemistry. A small number of *not publishable* sentences are thus to be expected.

Transl. method	Nr. of observations
MT	213
HT	216

Table 6.10: Number of edited incorrect sentences labelled as publishable.

Regarding *ok* sentences, the only condition where participants behaviour was correct beyond doubt is condition number 5 (see Table 6.9), since cases in which correct sentences were edited cannot always be considered as correct behaviours. Even if the final result was a publishable sentence, translators were not expected to modify correct sentences. Vice versa, when incorrect sentences were edited and then judged as improvable, the decision to edit the sentence was correct, but results were not. Also, if unedited correct sentences resulted in publishable ones, the possibility that this happened by chance cannot be ruled out.

The number of sentences for condition 5 in MT and in HT was compared to understand if for one of the HT or MT conditions a higher number of expected behaviours is observed.

Table 6.10 clearly shows that the number of sentences for condition 5 (see Table 6.9) is very similar for HT and MT, i.e. the number of behaviours that can undoubtedly be labelled as expected is the same for each translation method condition. This confirms what was also seen in Fig. 6.4, i.e. no warning signs were found that might lead one to think that revisers or post-editors behaviour was not always reliable because of flaws in the experiment structure.

After this sanity check, in the next sections data are analysed following the method described in Sect. 6.7.5.

6.9 Experiment results

6.9.1 HTER analysis

Tables 6.11 and 6.12 summarise significance and estimates for the effects of the two linear mixed models. As in Sect. 6.5.1, observations where HTER was higher than the mean plus 2 SD were removed. A total of 75 sentences were thus discarded. Figure 6.5 shows the distribution of HTER and WPS values for individual segments split by translation method and correctness.

As expected, in Figure 6.5 HTER is higher for incorrect sentences overall. While differences between post-editing and revision in both cases are small, HTER values for correct MT sentences are slightly higher than values for correct HT sentences. Comparing Figure 6.5 with Figure 6.2 highlights that in this final experiment students behaviour was more in line with the expectations. Boxes in Fig. 6.5 represents 50% of the observations for each conditions. For incorrect sentences – i.e. the ones where a higher HTER is expected – in the HT condition, 50% of the HTER values are included in the 13-30 HTER range. The median (the horizontal black line in the middle of the box) is slightly higher than 20. The same box in Fig. 6.2 has a bottom value close to 0, the highest value larger than 20 HTER, and the median is lower than 10 HTER. This shows that in this final experiment (Fig. 6.5) more edits were performed on incorrect HT sentences, while in the

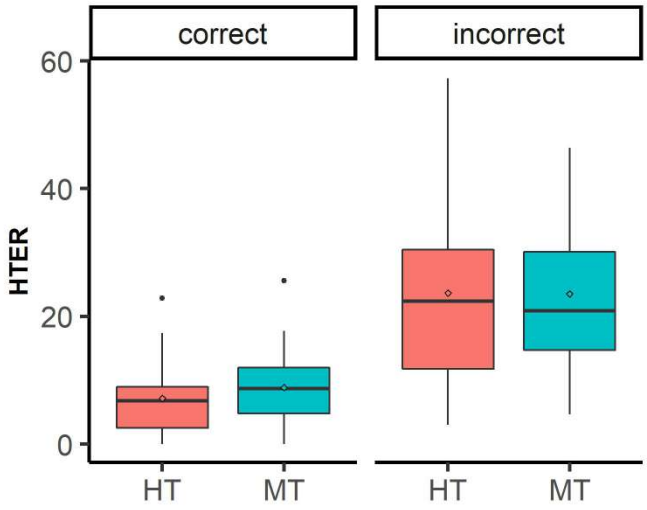


Figure 6.5: HTER values split by translation method and correctness of the translation.

HTER		
	Effect	***p* value**
Random	Participant	<0.001***
	Segment	<0.001***
Fixed	Correctness	<0.001***
	Method	0.6
	Interaction	0.14

Table 6.11: Significance of random and fixed effects on the two dependent variables: HTER.

Variables	**HTER**
HT incorrect	25.40
MT incorrect	23.60
HT correct	7.28
MT correct	8.99

Table 6.12: Estimates of the two linear mixed models for HTER. HTER goes up when more edits are performed.

pilot one (Fig. 6.2) many issues in incorrect sentences were not identified or solved (see Sect. 6.5.1 and 6.6), thus decreasing the HTER values.

Moving on to the results of the linear mixed model, the likelihood ratio test confirmed that the two random effects *participant* and *segment* do have a statistically significant impact on the HTER scores (see Table 6.11), i.e. the observations for the same segment or for the same participant are strongly correlated. Translation correctness is the only fixed effect with a statistically significant impact on HTER, while neither translation method nor its interaction with translation correctness significantly impact on it.

The model thus shows that the number of edits changes significantly only between correct and incorrect sentences, while the amount of edits performed on HT and MT sentences does not differ significantly. The effect of the interaction was not significant either, i.e. no significant change in HTER scores is observed in HT revision and post-editing across translation correctness conditions.

The similarity of the HTER values is confirmed by estimates in Table 6.12, where HTER is only slightly higher for MT correct sentences (+ 1.70), while the opposite happens in incorrect sentences, where HTER is higher for HT revised sentences (+1.71). Comparing these observations with those seen in Sect. 6.5.1, these look more in line with expectations since the gap between correct and incorrect sentences in terms of HTER is

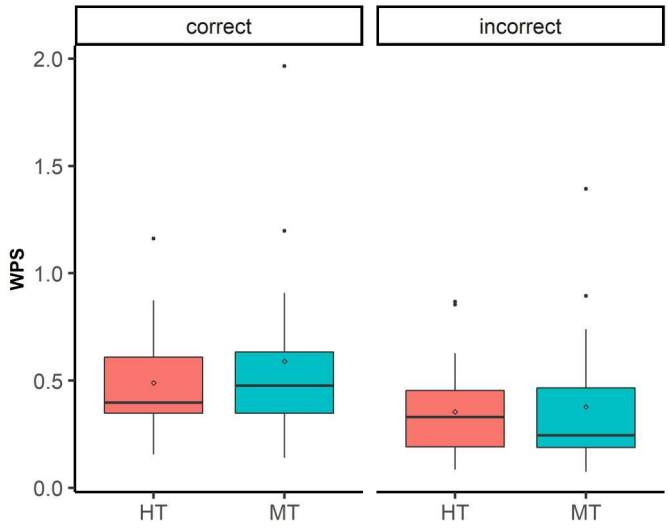


Figure 6.6: WPS values split by translation method and correctness of the translation.

larger.

Based on the results of the linear mixed model, HTER does not provide evidence of a lack of trust toward MT and proves that behaviours observed for both translation methods are similar.

6.9.2 WPS analysis

Also in this analysis outliers (observations with a WPS value higher than the mean plus 2 SDs) were removed, resulting in 152 discarded sentences.

Figure 6.6 shows that WPS is higher for correct sentences than for incorrect ones, while it is similar for PE and revision in the two conditions. As in Sect. 6.9.1, the p values in the WPS column of Table 6.13 confirm the statistically significant impact of the two random effects (participant and segment) on the dependent variable. However, in this case neither the two fixed effects (translation correctness and translation method), nor their interaction have a significant effect. This means that differences in terms of WPS between correct and incorrect sentences are not statistically significant. However, this should not be read as a warning sign, since as stated above (see Sect. 6.2 and 6.5.2) WPS also measures the cognitive effort. Moreover, the significant difference between correct and incorrect sentences in terms of HTER has already shown that participants' effort was higher on incorrect sentences. Since more edits do not necessarily require more time, it should not be expected that this same result is observed here. Similarly, significant differences between HT revision and PE were not found. When considering the interaction of translation method and translation correctness, WPS does not change significantly.

Looking at Table 6.14, participants were more productive on correct sentences than on incorrect ones, but values do not vary substantially. WPS is higher (+ 0.106) for correct MT sentences than for correct HT sentences, while for incorrect sentences productivity in terms of WPS is higher (+ 0.071) for HT than for MT.

Combining these results with those in Sect. 6.9.1 confirms that students did not trust

MT less than HT or vice versa.

WPS		
	Effect	***p* value**
Random	Participant	<0.001***
	Segment	<0.001***
Fixed	Correctness	0.118
	Method	0.436
	Interaction	0.433

Table 6.13: Significance of random and fixed effects on the two dependent variables: HTER.

Variables	**WPS**
HT incorrect	0.470
MT incorrect	0.399
HT correct	0.492
MT correct	0.598

Table 6.14: Estimates of the two linear mixed models for HTER. HTER goes up when more edits are performed.

6.9.3 Manual analysis

Differently from the pilot experiment, where a manual inspection was carried out to identify the reasons for unexpected behaviours, in this case a subset of the observations is analysed to understand if, even when differences between HT and MT are not found, some sentences reveal behaviours that could be further investigated in future work. To this aim, sentences with the largest differences in terms of mean HTER between MT and HT were extracted and examined.

Concerning Example 1 in Table 6.15, in both revision and PE, the same number of participants made the right decision, i.e. no edits. In the HT condition most of the participants who edited the sentence only changed the preposition. In the MT condition, terms were changed as well, resulting in a higher HTER mean score for MT (25.6) than for HT (17.3). Similarly in Example 2, most post-editors changed verb tenses or nominalised verbs. Mean HTER was 11.4 for MT and 6.79 for HT: most revisers did not edit the sentence.

Regarding incorrect sentences that were edited less in PE than revision, it would seem that revisers paid more attention to issues in the text than post-editors did. For example, all three occurrences of *reaction* in Example 3 should be plural and the term provided by the termbase is *Alkyl halides* rather than *Haloalkane*. 58.3% of the revisers spotted both issues, while only 34.78% of the post-editors did. As a result, mean HTER was 57.2 for HT revision and 43.4 for PE.

In Example 4, it would be sufficient to add the word *examination* at the end. However, in the HT condition most of the participants (54%) carried out a number of other edits applying to the whole sentence. Post-editors carried out unnecessary edits to a lesser extent (4.8%), such that mean HTER was 48.9 for HT and 43.8 for MT.

Slightly different behaviours are thus observed in sentences with a large difference in terms of mean HTER between MT and HT. However, evidence of one or more patterns characterising either post-editing or revision that could be further investigated in the rest of the data set or in future work was not found. Even if this analysis was only carried out on a subset of the sentences, it is possible to conclude that in this data set the qualitative analysis did not seem to provide insights diverging from those provided in the statistical analysis on temporal and technical effort.

Ex.	Sent. type	Text	Corr.
1	OUTPUT MT HT	Drugs during pregnancy, in children and in the elderly Drugs in children, in the elderly and during pregnancy Drugs during pregnancy, for children and for the elderly	Correct
2	OUTPUT MT HT	Finally, possible technical solutions to reduce the use of solvents and their recycling will be discussed Finally, possible technical solutions for solvent usage reduction and solvent recycling will be discussed Finally, possible technical solutions to reduce the use of solvents and to enable their recycling will be discussed.	Correct
3	OUTPUT MT HT	Haloalkane reactions (metal reaction, elimination reaction) Alkyl halides reactions (metal reaction, elimination reaction). Alkyl halides reactions (metal reactions, elimination reactions).	Incorrect
4	OUTPUT MT HT	The requirement to take the test is to have taken the Microbiology The requirement to take the test is to have taken the Microbiology examination. Only the students who passed the Microbiology test can take the exam.	Incorrect

Table 6.15: Examples of correct and incorrect outputs with large HTER differences between HT and MT.

6.9.4 Summing up

According to two analyses on HTER and WPS carried out through the use of linear models, significant changes were only found between HTER on correct and on incorrect sentences. No evidence of a lack of trust towards MT emerged. This behaviour confirms the positive opinion on MT stated in the pre-experiment questionnaire (see Table 6.8). This constructive attitude and the ability to interact with technology may be the result of greater awareness of the limits and strengths of MT and PE practice, acquired as part of their academic education (see Sect. 6.1 and 6.7.2).

While not significant, differences in terms of HTER and WPS do exist. Comparing Tables 6.12 and 6.14, in correct sentences both HTER and WPS are higher for MT than for HT, i.e. the larger the number of edits, the higher the productivity. In incorrect sentences, the lower the number of edits, the lower the productivity. These fluctuations are to be expected, since HTER is based on the number of edits, while WPS is also related to cognitive effort. High HTER scores are often linked to simple preferential changes (see Sect. 6.9.3), e.g. nominalisations and stylistic vocabulary variation. Such changes may be costly in terms of HTER, but do not require long searches or sentence restructuring – which would be costly in terms of WPS as well. If segments with complex terms are thoroughly checked with a focus on terminology, edits are less costly in terms of HTER than WPS, and discrepancies arise between WPS and HTER. Since participants are not expert in pharmacy or chemistry, terminology searches would not suggest distrust, while preferential changes would. Even though the manual analysis described above did not highlight particular editing patterns, it has to be noted that the presence of preferential changes in the edits might emerge if the whole data set were analysed. Future work might thus focus on a more thorough qualitative analysis, categorising the changes introduced in the different conditions and the attention-needing points in the raw output. A longer task would also be helpful, which would however increase fatigue and lead to possible adverse effects, especially since volunteer translator trainees are involved.

In Sect. 6.7.2 it was shown that students' professional experience is similar in both tasks, and that they are similarly acquainted with the basic notions of PE practice. Their familiarity with revision is certainly greater, though, as this is a standard component in

Question	Answers	MT part.	HT part.
Translation quality	Poor	0%	0%
	Sufficient	30.46%	60.87%
	Good	69.56%	39.13%
Termbase use	Not used	8.7%	4.35%
	Double checked	60.87%	73.91%
	Used	30.43%	21.74%

Table 6.16: Results of the post-experiment questionnaire on participants' perceived quality of the provided text and on their use of the termbase during the task.

translation courses at both BA and MA level. The more limited familiarity with PE might explain the WPS values obtained, which are highest for MT correct and lowest for MT incorrect. When a mistake is spotted in an MT-translated sentence, more time is spent choosing a strategy to edit it whereas, when a sentence is correct, it is quickly confirmed, as productivity is of the essence in PE. For HT revision, WPS results are more similar in both correctness conditions than is the case in MT. The lowest productivity observed in the MT incorrect condition would suggest that there is still scope for improving translators/post-editors trust in machine translation.

6.10 Post-experiment questionnaire

Trust is a multi-dimensional concept (see Chapter 4) and different methods should therefore be adopted to succesfully measure it. Since this was the first experiment on trust towards MT, it starts from the basic assumption that trust is strictly related to productivity (see Sect. 6.1), and thus measured HTER and WPS. To try and understand if students' behaviour and their perception of the text are correlated, a post-experiment questionnaire (see Appendix D) was prepared and students filled it in after carrying out the task. They were asked questions regarding the quality of the texts they worked on and regarding the use of the provided termbase.

Possible answers to the question on the quality of the provided text were:

- *Poor*. Almost every segment had to be edited and translation from scratch would have taken less time.
- *Sufficient*. Most of the segments had to be edited, but translating from scratch would have taken more time.
- *Good*. Only a few segments had to be edited and/or only minor changes were made.

Results are shown in Table 6.16. No participant rated text quality as *poor*, which is not surprising since sentences classified as completely wrong by the annotators were discarded from the text (see Sect. 6.4.2). What is interesting, however, is that the text quality was judged as *sufficient* by 60.87% of the revisers and by only 30.46% of the post-editors. Conversely, the proportion of *good* quality rating is much higher for post-editors (69.56%) than for revisers (39.13%). Results for MT confirm the positive attitude towards MT already seen in Table 6.8. On the other hand, the same quality level is only perceived as *sufficient* by most of the revisers because they probably tend to set higher expectations when they believe they are working on a text produced by humans.

The fact that these percentages do not correlate with differences between post-editors and revisers behaviour shows that there is a gap between participants' perception of the output quality and how they actually make use of the text. Participants probably lower their expectation regarding the text quality when post-editing, and this explains the percentages for the first question in Table 6.16. When they actually post-edit, their expectations regarding the text quality do not influence their behaviour, thus their effort is the same when post-editing and when revising even if perceived quality is higher for the text in the former condition. In this respect, it would have been interesting to ask students to evaluate each sentence before post-editing/revision. However, this task was not added to the experiment not to overload participants.

Another question was asked in the post-experiment questionnaire (see Appenidx D) and it concerned the use of the termbase provided to all participants (see Sect. 6.7.4). This question was asked starting from the assumption that participants not trusting the texts they were working on would have relied more on the termbase for terminology translation. When asked if they had used the termbase, participants could choose one of the following options:

- *Not used.* Participants did not look at the termbase.
- *Double checked.* Participants saw the termbase suggestions but double checked term translation when they were not sure about the translation.
- *Used.* Participants adopted the termbase suggestions when provided, without checking them further.

Table 6.16 shows percentages for each of these options. Only a small part of the participants stated that they had not used the termbase. One reviser stated that he/she already knew the translation of the terms provided in the termbase, for one MT participant the term translations provided in the output looked reliable enough and for the other MT participant a connection issue slowed down the upload of the termbase window. The number of participants who used the termbase without double checking is higher for MT (30.43%) than for HT (21.74%). 73.91% of the revisers used the termbase and then double checked the term, while only 60.87% of post-editors did the same. These numbers seem to suggest that post-editors relied slightly more on the termbase than revisers did (see Table 6.8). However, the total percentages of participants using the termbase, i.e. of participants stating that they used it or that they saw it and then double checked, was 95.65% for revisers and 91.3% for post-editors. Also, it has to be taken into account that one MT participant could not use the termbase due to technical issues, otherwise percentages could have been the same for the two translation methods. The choice of relying completely on the termbase might be more common for MT participants because in the module they attend on MT and post-editing they learn that terminology is one of the aspects MT still struggles the most with.

The answer to the two questions analysed here thus suggests that even though no significant differences between revisers and post-editors behaviour were found according to HTER and WPS, the perception regarding some parts of their task might change. When quality is in focus, the same text is deemed to be of a better quality if participants think it was produced with an MT system. Students thus tend to expect more in terms of quality from a text they believe to be produced by a human, while they lower their expectations when they know that the same text was machine translated.

Post-editors tend to rely on the provided aid without double checking slightly more than revisers. This seems to confirm what was introduced in Chapter 4, i.e. that in a situation of uncertainty a person tends to rely more on external aids, and for students post-editing is characterised by uncertainty more than revision, since they had little or no professional experience in MT and post-editing (see Table 6.8).

6.11 Conclusions and limitations

This chapter has described a study that investigated participants' trust towards MT compared to their trust towards HT. To this aim, a pilot experiment was first set up with participants with the same background as those taking part to the final experiment. Based on the limitations identified in the pilot experiment structure, a final experiment took place in March 2019. Starting from the assumption that trust is related to productivity (see Sect. 6.1), the analysis was based on HTER and WPS. A pre- and a post-experiment questionnaire were also filled in by students (see Appendices C and D).

While some flaws in the pilot experiment structure caused pilot participants' to behave in unexpected ways, in the final experiment the different statistical analyses proved that there is no difference between MT and HT in terms of participants' trust. Relating this result to those of the questionnaires, it became clear that students generally have a positive attitude towards MT – confirming results by Daems et al. (2017) and by Moorkens and O'Brien (2015) – which is reflected in their behaviour, since their lack of experience in post-editing does not affect their productivity with respect to revision. Also, they perceive the quality of a text to be higher when they think it is an MT output. Motivations for the positive attitude towards technologies can be found in the literature on trust (see Chapter 4) as well, i.e. when humans are in a situation of uncertainty they tend to rely more on provided aids, and post-editing is arguably a more uncertain scenario than revision for students.

This chapter has thus shown that there might be a new generation of translators that does not have preconceptions against MT. Reasons behind that are probably the background on MT and post-editing received during the Master's in Translation, the fact that this generation began studying translation after the neural outbreak – and thus with an increased output quality – and perhaps their lack of experience regarding the work conditions that are usually identified as the cause for a general reluctance of professional translators to work with MT (see Sect. 6.1). The possibility of carrying out the same experiment with professionals in the future would make it possible to understand how professional experience influences trust. While many studies have already focused on professional behaviour when post-editing (Cadwell et al., 2018; Sánchez-Gijón et al., 2019) or on the comparison of students and professionals (Daems et al., 2017; Moorkens and O'Brien, 2015), to the best of my knowledge, this is the first study focusing on investigating students' preconception against translation technology not with a view to projecting results on professional translators, but to understand students' attitude *per se*.

However, these observations and limitations should not hide the main finding of this study, namely that there are no significant differences between post-editors' and revisers' trust. This can be interpreted as a sign that, after receiving training on this new technology and before entering the translation industry, a new generation of translators does not seem to be affected by prejudice against MT as much as one could expect.

Chapter 7

Conclusion

7.1 Introduction

The present thesis has described an attempt at applying MT to institutional academic texts for Italian–English and German–English, specifically course catalogues. In order to reach this goal, two main research questions were investigated (see 1.1). The first one focused on the feasibility of profitably applying MT to such texts. Starting from the assumption that the benefits of a good quality MT are counteracted by possible preconceptions of translators towards the output, the second research question examined translators' trust towards an MT output *vs.* a HT.

Translating institutional academic texts raises two main issues (see Chapter 2). First, the scarcity of high-quality bilingual versions of course catalogues is a major bottleneck in the development of translation aids. When terminology is in focus, the high density of multi-domain terms – i.e. terms belonging to the education domain and terms belonging to a number of different disciplinary domains – and the lack of harmonisation are well-known issues. With a view to proposing solutions to these issues, the present work contributed to research in the MT field as follows. First, data sets were created for Italian–English and German–English in the institutional academic domains to train and test MT engines. In-depth evaluations of the output quality were provided for the engines in different scenarios using automatic metrics. Then, a gold standard with manual annotations of terminology was created and used for a terminology evaluation of the engines. The gold standard was released to stimulate research on terminology assessment. To conclude, trust was measured in a post-editing task and results regarding translator trainees' trust towards MT were outlined.

The aim of this final chapter is to summarise results related to each of the issues confronted in the present work, i.e. MT eligibility for institutional academic texts, terminology translation, and trust towards MT output. Implications of these results for future work and limitations are also discussed.

7.2 Results

Chapter 5 aimed at answering the research question "Can MT be profitably applied to the translation of institutional academic texts?". For both language combinations (It–En, De–En), degree programme and course unit descriptions were extracted from the websites of four universities. Given that texts from faculties whose discipline(s) belong to the macro-

category of the humanities were often not translated or difficult to extract, this macro-category was excluded. The resulting data sets were divided into a domain adaptation set (40,361 sentence pairs for It–En, 18,854 for De–En) and a test set (2,157 sentence pairs for It–En and 1,181 for De–En).

After the data collection step, two MT systems were tested in different scenarios using automatic metrics (BLEU and CharCut) to measure their performance. Both are SOTA systems and trained on a large pool of parallel data. ModernMT (MMT) was chosen because of its adaptation mechanism, which is able to leverage a domain adaptation data set to fine-tune a large pre-trained model on-the-fly. As a SOTA system, Google Translate (GT) provides an external validation of the quality of MMT. The first scenario represents an entry level in which universities want to start using MT to translate their course catalogues, but no bilingual data are available yet. Here, GT and MMT generic (henceforth MMT-I) are used. The second scenario is the one where universities start sharing their data and, after a few translation tasks, bilingual sentence pairs become available to be leveraged for domain adaptation. In this case MMT adapted (henceforth MMT-II) is used.

Results for the first scenario showed that it is possible to obtain an MT output of acceptable quality for both language combinations (GT achieves a BLEU score of 36.90 for Italian–English and 32.47 for German–English, MMT-I achieves 35.45 for Italian–English and 31.13 for German–English). In the second scenario, leveraging the relatively small data set for domain adaptation described above, an encouraging quality increase was observed (MMT-II achieves a BLEU score of 43.16 for Italian–English and 50.02 for German–English). Despite the lack of standardisation, domain adaptation can thus be leveraged to enhance the output quality. This is especially true for German–English, where the data set size was about half the size of the It–En one, and yet MMT-II brought a 18.89 increase in terms of BLEU with respect to MMT-I, while for It–En the increase was 7.71. This seems to show that the similarity between texts from the same institution is higher in the De–En data set than in the It–En one.

An additional scenario was tested, in which a new university – which has no parallel data yet – wants to start using MT, and bilingual course catalogues from other universities are leveraged. In this scenario results were disappointing, since domain adaptation did not bring any benefit in terms of quality. MMT-II was outperformed by GT for It–En (37.88 *vs.* 36.33 BLEU) and by MMT-I for De–En (35.95 *vs.* 35.69 BLEU). These results show that the lack of standardisation between texts produced by different universities hinders the possibility – for a university willing to start using MT – of leveraging data from other higher education institutions through the adaptation mechanism. However, the positive results obtained when no in-domain data are available (1st scenario), together with the quality increase obtained when a relatively small set of sentence pairs becomes available (2nd scenario), cast light on the feasibility of applying a static MT engine in the first translation tasks, resorting to the use of domain adaptation as soon as the first post-edited batches become available. In other words, the use of the most appropriate SOTA techniques, combined with CAT tools and the work of post-editors, is of the essence to overcome the lack of available bilingual data in the institutional academic domain. On the other hand, if MT starts to be consistently used by an increasing number of universities that agree to share their data, a virtuous circle could lead to increased standardisation and, as a consequence, to improved output quality and readability (see Sect. 2.2.3).

As stated above, the availability of bilingual course catalogues collected to train and test the MT engines also provided the opportunity for a focus on terminology trans-

lation. The test set – a collection of Italian–English sentence pairs extracted from 4 different universities – was annotated through a two-step manual process. Both single-word (SW) terms and multi-word (MW) ones appearing in the target sentence were first identified and categorised as `sure` – when their terminological status was certain – and `possible` – when their terminological status was not certain. Then, each annotated item was assigned to one of the following domains: `education` – including terms like *course*, *lecturer* –, `education equipment` – featuring education terms that are borrowed from other domains – e.g. *overhead projector* –, and `disciplinary` terms, i.e. terms belonging to the discipline taught in the course. The latter were then further split into specific disciplinary domains, ranging from biology to chemistry and electrical engineering. The resulting data set features 7,517 terms (3,916 SWs and 3,601 MWs) distributed in 2,055 sentence pairs and 22 domains, two of them being `education` and `education equipment`, and the other 20 being different `disciplinary` domains. 17 disciplinary domains out of 20 contain more than 100 terms.

The gold standard was then used to assess how the three engines used in this work (GT, MMT-I and MMT-II) handle multi-domain terminology. To this aim, the Term Hit Rate (THR) metric (Farajian et al., 2018) was used to compute the proportion between the number of terms correctly generated by the MT system and the total number of terms annotated in the gold standard (i.e. that should have been generated by the system). Results showed that terms belonging to the `education` and `education-equipment` domains are the most difficult to translate for MT systems, although domain adaptation brings an increase in the THR score. MMT-II outperformed MMT-I by 5.97 THR on `education` terms and by 3.33 on `equipment` terms, and it outperformed GT by more than 3 THR on both `education` and `equipment` terms. Nonetheless, THR for these two domains remained lower than in all other domains. MWs are also complex to handle for NMT. For example, the best THR on the total number of SWs was 76.07 for MMT-II, while for MWs the best performing system was again MMT-II, but with a 53.65 THR. This trend was confirmed for MWs in the `disciplinary`, `education` and `equipment` domains. The highest THR scores were observed in the `disciplinary` category (67.74 for MMT-II on the whole amount of disciplinary terms). Since it is uncommon for disciplinary concepts to have more than one term referring to them, terminology in this domain is likely to be standardised and uniform, while in the `education` and `equipment` domains, THR scores are lower because of the aforementioned lack of standardisation. As a matter of fact, the 5 most frequent MWs in the data set include: *oral exam, written test, oral examination*, i.e. 3 different terms to refer to the concept of *exam*. Once again, results showed that striving for more uniformity across texts from different universities would make it easier to develop translation aids in this domain. In this case as well, using MT and CAT tools consistently could bring considerable improvements to the translation pipeline, not to mention text quality and readability.

The gold standard, called MAGMATic, was released with the aim of stimulating further research on terminology evaluation[1]. As a matter of fact, to the best of my knowledge NMT evaluations so far were focused on lexical issues rather than concentrating specifically on terminology (Bentivogli et al., 2016, 2018; Toral and Sánchez-Cartagena, 2017; Van Brussel et al., 2018).

Summing up this first part, the present work has shown that MT can be used to translate institutional academic texts with a satisfactory quality. This is especially true if more

[1]MAGMATic is freely downloadable under CC BY–NC–SA 4.0 from: `https://ict.fbk.eu/magmatic/`

consistent bilingual data become available for a number of universities.

The second research question addressed in this work was "Do translators trust MT?". 1st year students from a Master's in Translation took part in an experiment where their trust towards MT was compared to their trust towards HT. The choice of carrying out this study with students is motivated by a number factors. First, today's translation students will be among the first translators to enter the market with a background on MT and post-editing, since an increasing number of degrees in translation are starting to offer these courses. Also, becoming familiar with MT after the beginning of the neural era might cause students to be more willing to trust MT given the increase in the output quality. With respect to working with professionals, variables related to the years of experience and/or knowledge of the languages are kept under control working with students from the same cohort (see Sect. 6.4.1 and 6.7.2).

After a pilot experiment that took place in May 2018 and made it possible to identify some flaws in the structure, a final experiment was carried out in March 2019. All participants were given the same target text. Half of them were told it was a HT needing revision, half of them that it was an MT output to be post-edited. Both groups were asked to provide a final version of a publishable quality for the website of a university. Two linear mixed models were built to measure differences between MT and HT in terms of number of edits (HTER was computed between the raw output and the edited versions) and of temporal effort (words per second, WPS). Results showed that participants' editing and temporal effort in MT and HT do not differ significantly. A manual inspection of the sentences with the largest differences between MT and HT in terms of mean HTER proved the absence of preferential changes in either MT or HT. The data set and the statistical analyses results thus clearly revealed that students had the same level of trust in both MT and HT. This was supported by a pre-experiment questionnaire, where the vast majority of the participants declared to have a positive or very positive opinion on MT and its usefulness for translators. Interestingly, in a post-experiment questionnaire post-editors tended to rate the output quality as *very good*, while the majority of those working on the purported HT task rated the same text as *sufficiently good*. This showed that, despite their positive opinion on MT and their behaviour when post-editing, students still tend to lower their expectations regarding the text quality when they think or know it is the output of a machine. Going back to the main finding of this study, the absence of a preconception of students against MT might be the proof of what was postulated above, i.e. that receiving training on MT and post-editing, and becoming familiar with this technology after the quality increase brought by neural networks, can have an impact on translator trainees preconceptions against MT. At the same time, students are arguably not influenced by the downward trend of translation rates brought by MT post-editing, which might often be the cause for a reluctance to use MT among professionals.

To answer this second research question, the new generation of translators trust the MT output the same way they trust a HT. Considering what was posited in Sect. 1.1, this result also confirms that the quality obtained from the engines described in Chapter 5 is not counterbalanced by a reluctance of post-editors to use the output.

7.3 Limitations and future work

Given the positive results obtained in the present work, and considering the novelty of the research questions posed here, future work might fruitfully deal with MT and institutional academic texts, but also with terminology evaluation and trust assessment. Possible future

work might be useful to overcome the limitations of the present one as well.

Regarding the application of MT to the institutional academic domain, the focus was on two initial scenarios, i.e. one where no data is available and one where a relatively small amount of data becomes available for a number of universities. A further step might be that of adding more data for one university only, i.e. a scenario in which one university in particular translates a good amount of texts that can be used for domain adaptation. This could provide information on possible changes in the output quality when texts from that same university and from other universities as well are translated. In general, a data augmentation step – either with data from one university only or with data from several ones – could also be followed by new tests on the additional scenario, i.e. the one where data from several universities are leveraged to translate texts from a university that has no bilingual data yet. Since the first tests in this additional scenario showed that texts produced by different universities diverge too much for the adaptation mechanism to be beneficial, it would be interesting to evaluate the effects of an increase in the size of the domain adaptation data set.

If a number of universities decided to start combining MT and CAT tools in their translation pipeline and to share their data, the data set would increase substantially. This could provide the opportunity to train a dedicated engine and compare advantages and disadvantages of adapting a large pre-trained model *vs.* training a model from scratch with in-domain data only.

Regarding the part of the work on terminology, MAGMATic could be used to assess term translation after the data augmentation step(s) mentioned above. While THR only looked for the occurrence of reference terms in the output, a manual evaluation of the whole data set might be useful to understand how terms were placed in context. Besides, THR was computed on the exact match of reference terms. Adding a lemmatisation step before the assessment, it would be possible to measure the impact of morphology on the number of correctly translated reference terms.

Not having source annotations might be seen as a limitation of the present work on terminology. However, it has to be noted that annotating the source as well would have added an extra step to an already demanding task, increasing the risk of a fatigue effect for annotators. Also, the present contribution has shown that accurate terminology assessment is possible with target annotation only. Adding more language combinations to the data set might increase its usefulness. Therefore, future work might focus on building the same data set for German–English, since institutional academic bilingual sentences are already available for this language combination. Furthermore, it has to be noted that the release of the MAGMATic data set can be extremely useful for the MT community, since it allows for in-depth terminology assessments of MT systems. The release can stimulate similar work in other domains or fields as well, especially thanks to the multi-domain nature of the data set.

In the chapter on trust assessment it was suggested that the new generation of translators does not appear to have prejudices against MT. Reasons why students were chosen for this experiment were mentioned in Sect. 7.2. Nonetheless, involving both professionals and students would allow for a thorough comparison of their level of trust, perhaps revealing if professionals do show a less positive attitude than students, how possible differences reflect on the practical task and how they originated. Indeed, collecting observations and feedback from professionals would allow to have an insight on how the years of experience and the positive or negative market trends impact on preconceptions and productivity.

Seeing trust as strictly related to productivity seemed to be the ideal starting point in this first attempt at measuring trust during a real task. However, more studies would be needed to shed light on the complex and multi-dimensional nature of trust. In both HT and MT conditions, participants worked on an MT output composed of correct and incorrect sentences. Another interesting condition would be to provide participants with a post-edited/revised version of a raw output, and ask them to post-edit/revise the text. If they tend to modify it anyway – and if the number of edits is higher when they think they are revising *vs.* post-editing – it would mean that they have a preconception against MT or HT. Also, an experiment where participants post-edit sentences produced by different MT systems might help investigating possible preconceptions against, e.g., a system they are not familiar with. Pre- and post-experiment interviews could better clarify what participants expect from a HT *vs.* an MT output, and why. As a matter of fact, while productivity and number of edits did not reveal differences between HT and MT, a closer look at participants' perception – e.g. through post-experiment interviews or more extended questionnaires – might help to spot interesting patterns between perceived quality and actual behaviour along the lines of Sánchez-Gijón et al. (2019). To conclude, it has to be taken into account that – especially when working with students without professional experience – keeping the task as simple as possible is of utmost importance, to make sure that their performance is not influenced by the duration of the experiment or the number of assignments.

From a general point of view, the work on MT and institutional academic texts has shown the importance of sharing data and translation practices to achieve better performance. If a large number of universities agreed to apply MT to their course catalogues, to use CAT tools to improve the efficiency of the post-editing process, and to share the resulting bilingual texts, quality would probably increase in a relatively short span of time, streamlining the translation process of all universities involved. Moreover, the focus on terminology has underlined the importance of involving linguists in the development of data sets that can provide a more insightful assessment of the output.

To conclude, the experiment on trust emphasised the role of training to prevent or limit a possible lack of trust. This result is particularly relevant for modern society and work environments. An increasing number of tasks are now carried out with the help of technologies – and artificial intelligence in particular –, in many cases without providing users with the appropriate knowledge of their main characteristics. This might prevent a fruitful human-machine interaction, while being aware of the strengths and weaknesses of technologies – as in the experiment described above – helps users to establish a beneficial relationship with them.

Bibliography

Alves, Fabio and Daniel Vale (2009). "Probing the unit of translation in time: Aspects of the design and development of a web application for storing, annotating, and querying translation process data". In: *Across Languages and Cultures* 10.2, pp. 251–273. DOI: 10.1556/Acr.10.2009.2.5. eprint: https://doi.org/10.1556/Acr.10.2009.2.5. URL: https://doi.org/10.1556/Acr.10.2009.2.5.

Arcan, Mihael and Paul Buitelaar (2017). "Translating Domain-Specific Expressions in Knowledge Bases with Neural Machine Translation". In: *CoRR* abs/1709.02184. arXiv: 1709.02184. URL: http://arxiv.org/abs/1709.02184.

Arcan, Mihael, Marco Turchi, Sara Tonelli, and Paul Buitelaar (2014a). "Enhancing Statistical Machine Translation with Bilingual Terminology in a CAT Environment". In: *Proceedings of AMTA 2014*. Ed. by Yaser Al-Onaizan and Michel Simard. Vancouver, BC.

Arcan, Mihael, Claudio Giuliano, Marco Turchi, and Paul Buitelaar (2014b). "Identification of Bilingual Terms from Monolingual Documents for Statistical Machine Translation". In: *Proceedings of the 4th International Workshop on Computational Terminology*. Dublin, Ireland, pp. 22–31. URL: http://www.aclweb.org/anthology/W14-4803.

Arnold, D., L. Balkan, R.L. Humphreys, S. Meijer, and L. Sadler (1994). *Machine Translation: an Introductory Guide*. London: NCC Blackwell. URL: http://www.essex.ac.uk/linguistics/external/clmt/MTbook/PostScript/.

Arthur, Philip, Graham Neubig, and Satoshi Nakamura (2016). "Incorporating Discrete Translation Lexicons into Neural Machine Translation". In: *CoRR* abs/1606.02006. arXiv: 1606.02006. URL: http://arxiv.org/abs/1606.02006.

Artstein, Ron and Massimo Poesio (2008). "Inter-coder Agreement for Computational Linguistics". In: *Comput. Linguist.* 34.4, pp. 555–596. ISSN: 0891-2017. DOI: 10.1162/coli.07-034-R2. URL: http://dx.doi.org/10.1162/coli.07-034-R2.

Astrakhantsev, Nikita A., Denis G. Fedorenko, and Denis Yu. Turdakov (2015). "Methods for automatic term recognition in domain-specific text collections: A survey". In: *Programming and Computer Software* 41.6, pp. 336–349. ISSN: 1608-3261. DOI: 10.1134/S036176881506002X. URL: https://doi.org/10.1134/S036176881506002X.

Ba, Sulin and Paul A. Pavlou (2002). "Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior". In: *MIS Q.* 26.3, pp. 243–268. ISSN: 0276-7783. DOI: 10.2307/4132332. URL: http://dx.doi.org/10.2307/4132332.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473. URL: http://arxiv.org/abs/1409.0473.

Baker, Paul (2010). *Sociolinguistics and Corpus Linguistics*. English. Edinburgh University Press, pp. 19–21. ISBN: 9780748627356.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico (2016). "Neural versus Phrase-Based Machine Translation Quality: a Case Study". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 257–267. DOI: 10.18653/v1/D16-1025. URL: http://www.aclweb.org/anthology/D16-1025.

— (2018). "Neural versus phrase-based MT quality: An in-depth analysis on English-German and English-French". In: *Computer Speech & Language* 49, pp. 52–70. DOI: 10.1016/j.csl.2017.11.004. URL: https://doi.org/10.1016/j.csl.2017.11.004.

Bernier-Colborne, Gabriel and Patrick Drouin (2014). "Creating a test corpus for term extractors through term annotation". In: *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 20.1, pp. 50–73. ISSN: 0929-9971. DOI: https://doi.org/10.1075/term.20.1.03ber. URL: https://www.jbe-platform.com/content/journals/10.1075/term.20.1.03ber.

Bertoldi, Nicola, Mauro Cettolo, and Marcello Federico (2013). "Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation". In: *Proceedings of the XIV Machine Translation Summit*. Ed. by Andy Way, Khalil Sima'an, Mikel L. Forcada, Daniel Grasmick, and Heidi Depaetere. Nice, France, pp. 35–42.

Bertoldi, Nicola et al. (2017). "MMT: New Open Source MT for the Translation Industry". In:

Bertoldi, Nicola, Davide Caroselli, and Marcello Federico (2018). "The ModernMT Project". In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*. Alacant, Spain.

Bisazza, Arianna, Nick Ruiz, and Marcello Federico (2011). "Fill-up versus interpolation methods for phrase-based SMT adaptation". In: *2011 International Workshop on Spoken Language Translation, IWSLT 2011, San Francisco, CA, USA, December 8-9, 2011*, pp. 136–143.

Blomqvist, Kirsimarja (1997). "The many faces of trust". In: *Scandinavian Journal of Management* 13.3, pp. 271 –286. ISSN: 0956-5221. DOI: https://doi.org/10.1016/S0956-5221(97)84644-1. URL: http://www.sciencedirect.com/science/article/pii/S0956522197846441.

Bojar, Ondřej et al. (2016). "Findings of the 2016 Conference on Machine Translation". In: *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, pp. 131–198. URL: http://www.aclweb.org/anthology/W/W16/W16-2301.

Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum (2012). "Identifying bilingual Multi-Word Expressions for Statistical Machine Translation". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: European Languages Resources Association (ELRA), pp. 674–679. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/886_Paper.pdf.

Cabré, M.T. and J.C. Sager (1999). *Terminology: Theory, Methods, and Applications*. Terminology and lexicography research and practice. J. Benjamins Publishing Com-

pany. ISBN: 9789027216342. URL: https://books.google.it/books?id=GAqGD9Xtu0IC.

Cadwell, Patrick, Sharon O'Brien, and Carlos S. C. Teixeira (2018). "Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators". In: *Perspectives* 26.3, pp. 301–321. DOI: 10.1080/0907676X.2017.1337210. eprint: https://doi.org/10.1080/0907676X.2017.1337210. URL: https://doi.org/10.1080/0907676X.2017.1337210.

Callahan, Ewa and Susan C. Herring (2012). "Language choice on university websites: Longitudinal trends". In: *Journal of International Communication* 6 (2012), pp. 322–355.

Candel-Mora, Miguel Ángel and María Luisa Carrió-Pastor (2014). "Terminology Standardization Strategies towards the Consolidation of the European Higher Education Area". In: *Procedia - Social and Behavioral Sciences* 116, pp. 166 –171.

Carl, Michael (2012). "Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research." In: *LREC*. Ed. by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. European Language Resources Association (ELRA), pp. 4108–4112. ISBN: 978-2-9517408-7-7. URL: http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html#Carl12.

Castilho, Sheila, Federico Gaspari, Joss Moorkens, and Andy Way (2017a). "Integrating machine translation into MOOCs". In: *International Conference on Education and New Learning Technologies*. Ed. by Luis Gómez Chova, Agustín López Martínez, and Ignacio Candel Torres. Vol. 1. Barcelona, Spain: IATED, pp. 9360 –9365. ISBN: 978-84-697-3777-4. DOI: 10.21125/edulearn.2017.0765. URL: https://www.researchgate.net/publication/318706512_INTEGRATING_MACHINE_TRANSLATION_INTO_MOOCS.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way (2017b). "Is Neural Machine Translation the New State of the Art?" In: *The Prague Bulletin of Mathematical Linguistics* 108.1. Exported from https://app.dimensions.ai on 2018/09/13, pp. 109–120. DOI: 10.1515/pralin-2017-0013. URL: https://app.dimensions.ai/details/publication/pub.1085921590andhttp://www.degruyter.com/downloadpdf/j/pralin.2017.108.issue-1/pralin-2017-0013/pralin-2017-0013.xml.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Andy Way, and Panayota Georgakopoulou (2018). "Evaluating MT for massive open online courses". In: *Machine Translation*. ISSN: 1573-0573. DOI: 10.1007/s10590-018-9221-y. URL: https://doi.org/10.1007/s10590-018-9221-y.

Chatterjee, Rajen, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain (2017). "Guiding Neural Machine Translation Decoding with External Knowledge". In: *Proceedings of the Conference on Machine Translation (WMT)*. Vol. 1. Copenhagen, Denmark: Association for Computational Linguistics, pp. 157–168.

Cho, Kyunghyun, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Pro-*

cessing (EMNLP). Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. URL: http://www.aclweb.org/anthology/D14-1179.

Crosier, David, Lewis Purser, and Hanne Smidt (2007). *Universities Shaping the EHEA*. Brussels: European University Association.

Daems, Joke, Sonia Vandepitte, Robert Hartsuiker, and Lieve Macken (2017). "Translation methods and experience : a comparative analysis of human translation and post-editing with students and professional translators". eng. In: *META* 62.2, pp. 245–270. ISSN: 0026-0452. URL: http://dx.doi.org/10.7202/1041023ar.

Depraetere, Heidi, Joachim Van den Bogaert, and Joeri Van de Walle (2011). "Bologna Translation Service: Online translation of course syllabi and study programmes in English". In: *Proceedings of the 15th Conference of the European Association for Machine Translation*. Ed. by Mikel L. Forcada, Heidi Depraetere, and Vincent Vandeghinste. Leuven, Belgium, pp. 29–34.

Dice, Lee Raymond (1945). "Measures of the Amount of Ecologic Association Between Species". In: *Ecology* 26.3, pp. 297–302. URL: http://www.jstor.org/pss/1932409.

Doddington, George (2002). "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics". In: *Proceedings of the Second International Conference on Human Language Technology Research*. HLT '02. San Diego, California: Morgan Kaufmann Publishers Inc., pp. 138–145. URL: http://dl.acm.org/citation.cfm?id=1289189.1289273.

Farajian, M. Amin, Marco Turchi, Matteo Negri, and Marcello Federico (2017). "Multi-Domain Neural Machine Translation through Unsupervised Adaptation". In: *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 127–137. DOI: 10.18653/v1/W17-4713. URL: http://aclweb.org/anthology/W17-4713.

Farajian, M. Amin, Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico (2018). "Evaluation of Terminology Translation in Instance-Based Neural MT Adaptation". In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. Alacant, Spain.

Fernandez Costales, Alberto (2012). "The internationalization of institutional websites". In: *Translation Research Projects*. Ed. by Anthony Pym and David Orrego-Carmona. Tarragona: Intercultural Studies Group, pp. 51–60.

Ferraresi, Adriano (2016). *Collocations in institutional academic English. Corpus and experimental perspectives*. ISBN: 978-88-548-9393-1.

— (2017). "Terminology in European University Settings. The Case of Course Unit Descriptions". In: *Terminological Approaches in the European Context*. Ed. by Paola Faini. Cambridge Scholars Publishing, Newcastle upon Tyne, pp. 20–40.

Ferraresi, Adriano and Silvia Bernardini (2013). "The academic Web-as-Corpus". In: *8th Web a k8ck s Corpus workshop (WAC-8)*. Lancaster, UK.

Flanagan, Mary A. (1994). "Error classification for mt evaluation". In: *In Proceedings of 1st Conference of the Association for Machine Translation in the Americas (AMTA*, pp. 65–72.

Girardi, Christian, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico (2014). "MT-EQuAl: a Toolkit for Human Assessment of Machine Translation Output". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin, Ireland: Dublin City University

and Association for Computational Linguistics, pp. 120–123. URL: https://www.aclweb.org/anthology/C14-2026.

Glanville, Jennifer L. and Pamela Paxton (2007). "How do We Learn to Trust? A Confirmatory Tetrad Analysis of the Sources of Generalized Trust". In: *Social Psychology Quarterly* 70.3, pp. 230–242. DOI: 10.1177/019027250707000303. eprint: https://doi.org/10.1177/019027250707000303. URL: https://doi.org/10.1177/019027250707000303.

Gries, Th. Stefan (2015). "The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models". In: *Corpora* 10.1, pp. 95–125. DOI: 10.3366/cor.2015.0068. eprint: https://doi.org/10.3366/cor.2015.0068. URL: https://doi.org/10.3366/cor.2015.0068.

Hasler, Eva, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne (2018). "Neural Machine Translation Decoding with Terminology Constraints". In: *CoRR* abs/1805.03750. arXiv: 1805.03750. URL: http://arxiv.org/abs/1805.03750.

Hassan, Hany et al. (2018). "Achieving Human Parity on Automatic Chinese to English News Translation". In: *CoRR* abs/1803.05567. arXiv: 1803.05567. URL: http://arxiv.org/abs/1803.05567.

Hutchins, W. John (1995). "Machine Translation: A Brief History". In: *Concise history of the language sciences: from the Sumerians to the cognitivists, Pergamon*. Press, pp. 431–445.

— (2007). *Machine translation: A concise history*.

Jalili Sabet, Masoud, Matteo Negri, Marco Turchi, José GC de Souza, and Marcello Federico (2016). "TMop: a Tool for Unsupervised Translation Memory Cleaning". In: pp. 49–54.

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). "GENIA corpus—a semantically annotated corpus for bio-textmining". In: *Bioinformatics* 19.suppl_1, i180–i182. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btg1023. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/19/suppl_1/i180/614820/btg1023.pdf. URL: https://doi.org/10.1093/bioinformatics/btg1023.

Koehn, Philipp (2010). *Statistical Machine Translation*. 1st. New York, NY, USA: Cambridge University Press. ISBN: 0521874157, 9780521874151.

Koehn, Philipp and Rebecca Knowles (2017). "Six Challenges for Neural Machine Translation". In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver: Association for Computational Linguistics, pp. 28–39. DOI: 10.18653/v1/W17-3204. URL: https://www.aclweb.org/anthology/W17-3204.

Koehn, Philipp et al. (2007). "Moses: Open Source Toolkit for Statistical Machine Translation". In: *Proceedings of the ACL 2007 Demo and Poster Sessions*. Prague: Association for Computational Linguistics, pp. 177–180.

Krings, Hans P. (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent, Ohio: Kent State University Press.

Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1.

Langlais, Philippe (2002). "Improving a General-purpose Statistical Translation Engine by Terminological Lexicons". In: *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14*. COMPUTERM '02. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 1–7. DOI:

10.3115/1118771.1118776. URL: http://dx.doi.org/10.3115/1118771.1118776.

Lardilleux, Adrien and Yves Lepage (2017). "CHARCUT: Human-Targeted Character-Based MT Evaluation with Loose Differences". In: *Proceedings of IWSLT 2017*. Tokyo, Japan. URL: https://hal.archives-ouvertes.fr/hal-01726326.

Läubli, Samuel and David Orrego-Carmona (2017). "When Google Translate is better than Some Human Colleagues, those People are no longer Colleagues". In: *Translating and the Computer 39*. London, pp. 56–69.

Lee, John D. and Katrina A. See (2004). "Trust in automation: designing for appropriate reliance". In: *Human Factors* 46.1, pp. 50–80. URL: http://www.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf.

Lommel, Arle, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit (2014). "Using a new analytic measure for the annotation and analysis of MT errors on real data". In: *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. Dubrovnik, Croatia, pp. 165–172.

Luhmann, Niklas (2000). "Familiarity, Confidence, Trust: Problems and Alternatives". In: *Trust: Making and Breaking Cooperative Relations*. Ed. by Diego Gambetta. electronic. Department of Sociology, University of Oxford. Chap. 6, pp. 94–107.

Luong, Minh-Thang and Christopher D. Manning (2015). "Neural Machine Translation Systems for Spoken Language Domains". In: *Proceedings of IWSLT*. Da Nang, Vietnam.

Luong, Thang, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba (2015). "Addressing the Rare Word Problem in Neural Machine Translation". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pp. 11–19. URL: http://aclweb.org/anthology/P/P15/P15-1002.pdf.

Madhavan, Poornima and Douglas A. Wiegmann (2007). "Similarities and differences between human–human and human–automation trust: an integrative review". In: *Theoretical Issues in Ergonomics Science* 8.4, pp. 277–301. DOI: 10.1080/14639220500337708. eprint: https://doi.org/10.1080/14639220500337708. URL: https://doi.org/10.1080/14639220500337708.

Martindale, Marianna J. and Marine Carpuat (2018). "Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT". In: abs/1802.06041. arXiv: 1802.06041. URL: http://arxiv.org/abs/1802.06041.

McKnight, D. Harrison and Norman L. Chervany (2001). "What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology". In: *International Journal of Electronic Commerce* 6.2, pp. 35–59. ISSN: 1086-4415. DOI: 10.1080/10864415.2001.11044235. URL: https://doi.org/10.1080/10864415.2001.11044235.

Moorkens, Joss and Sharon O'Brien (2015). "Post-Editing Evaluations: Trade-offs between Novice and Professional Participants". In: *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*. Antalya, Turkey, pp. 75–81. URL: https://www.aclweb.org/anthology/W15-4910.

Moorkens, Joss, Sharon O'Brien, Igor A. L. da Silva, Norma B. de Lima Fonseca, and Fábio Alves (2015). "Correlations of perceived post-editing effort with measurements of actual effort." In: *Machine Translation* 29.3-4, pp. 267–284.

Nagao, M. (1984). "A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle". In: *Artificial and Human Intelligence*. Ed. by A. Elithorn and R. Barnerji. North-Holland, pp. 173–180.

Och, Franz Josef (2003). "Minimum Error Rate Training in Statistical Machine Translation". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*. ACL '03. Sapporo, Japan: Association for Computational Linguistics, pp. 160–167. DOI: 10.3115/1075096.1075117. URL: https://doi.org/10.3115/1075096.1075117.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: A Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://doi.org/10.3115/1073083.1073135.

Pinnis, Mārcis and Raivis Skadinš (2012). "MT Adaptation for Under-Resourced Domains - What Works and What Not". In: *Human Language Technologies - The Baltic Perspective - Proceedings of the Fifth International Conference Baltic HLT 2012*. Tartu, Estonia, pp. 176–184. DOI: 10.3233/978-1-61499-133-5-176. URL: https://doi.org/10.3233/978-1-61499-133-5-176.

Pokorn, Nike K. (2016). "Is it so different? Competences of teachers and students in L2 translation classes". In: *International Journal of Translation* 18, pp. 31–48. ISSN: 1722-5906. DOI: 10.13137/2421-6763/13664.

Popović, Maja (2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049. URL: https://www.aclweb.org/anthology/W15-3049.

Pu, Xiao, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis (2018). "Integrating Weakly Supervised Word Sense Disambiguation into Neural Machine Translation". In: *CoRR* abs/1810.02614. arXiv: 1810.02614. URL: http://arxiv.org/abs/1810.02614.

Q. Zadeh, Behrang and Siegfried Handschuh (2014). "The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics". In: *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, pp. 52–63. DOI: 10.3115/v1/W14-4807. URL: https://www.aclweb.org/anthology/W14-4807.

Rempel, John K., John G. Holmes, and Mark P. Zanna (2004). "Trust in Close Relationships". In:

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang (2009). "Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions". In: *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. MWE '09. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 47–54. ISBN: 978-1-932432-60-2.

Rossetti, Alessandra and Federico Gaspari (2017). "Modelling the Analysis of Translation Memory Use and Post-Editing of Raw Machine Translation Output: A Pilot Study of Trainee Translators' Perceptions of Difficulty and Time Effectiveness." In: *Empirical Modelling of Translation and Interpreting*. Ed. by Silvia Hansen-Schirra, Oliver Czulo, and Hofmann Sascha. Berlin: Language Science Press., pp. 41–67. DOI: doi:10.5281/zenodo.1090952.

Sánchez-Gijón, Pilar, Joss Moorkens, and Andy Way (2019). "Post-editing neural machine translation versus translation memory segments". In: *Machine Translation* 33.1-2, pp. 31–59. DOI: 10.1007/s10590-019-09232-x. URL: https://doi.org/10.1007/s10590-019-09232-x.

Scansani, Randy, Marcello Federico, and Luisa Bentivogli (2017a). "Assessing the Use of Terminology in Phrase-Based Statistical Machine Translation for Academic Course Catalogues Translation". In: *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*. URL: http://ceur-ws.org/Vol-2006/paper074.pdf.

Scansani, Randy, Silvia Bernardini, Adriano Ferraresi, Federico Gaspari, and Marcello Soffritti (2017b). "Enhancing Machine Translation of Academic Course Catalogues with Terminological Resources". In: *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*. Varna, Bulgaria: Association for Computational Linguistics, Shoumen, Bulgaria, pp. 1–10. DOI: 10.26615/978-954-452-042-7_001.

Scansani, Randy, Silvia Bernardini, Adriano Ferraresi, and Luisa Bentivogli (2019a). "Do translator trainees trust machine translation? An experiment on post-editing and revision". In: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation, pp. 73–79. URL: https://www.aclweb.org/anthology/W19-6711.

Scansani, Randy, Luisa Bentivogli, Silvia Bernardini, and Adriano Ferraresi (2019b). "MAGMATic: A Multi-domain Academic Gold Standard with Manual Annotation of Terminology for Machine Translation Evaluation". In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland: European Association for Machine Translation, pp. 78–86. URL: https://www.aclweb.org/anthology/W19-6608.

Sennrich, Rico and Barry Haddow (2016). "Linguistic Input Features Improve Neural Machine Translation". In: *CoRR* abs/1606.02892. arXiv: 1606.02892. URL: http://arxiv.org/abs/1606.02892.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://www.aclweb.org/anthology/P16-1162.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006). "A study of translation edit rate with targeted human annotation". In: *Proceedings of Association for Machine Translation in the Americas*. Cambridge, Massachussets, pp. 223–231.

Somers, Harold (1992). *An Introduction to Machine Translation*. London Academic Press.

Specia, Lucia, Najeh Hajlaoui, Catalina Hallett, and Wilker Aziz (2011). "Predicting Machine Translation Adequacy". In: *Proceedings of the 13th Machine Translation Summit*. Xiamen, China, pp. 513–520.

Štajner, Sanja, Andreia Querido, Nuno Rendeiro, João António Rodrigues, and António Branco (2016). "Use of Domain-Specific Language Resources in Machine Translation". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 592–598. ISBN: 978-2-9517408-9-1.

Stanojević, Miloš and Khalil Sima'an (2015). "BEER 1.1: ILLC UvA submission to metrics and tuning task". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 396–401. DOI: 10.18653/v1/W15-3050. URL: https://www.aclweb.org/anthology/W15-3050.

Stewart, Dominic (2000). "Poor Relations and Black Sheep in Translation Studies". In: *Target: International Journal of Translation Studies* 12.02, pp. 205–228.

— (2011). "Translation textbooks: translation into English as a foreign language". In: *inTRAlinea Online Translation Journal*. URL: http://www.intralinea.org/review_articles/article/1541.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, pp. 3104–3112. URL: http://dl.acm.org/citation.cfm?id=2969033.2969173.

Tezcan, Arda, VÃ©ronique Hoste, and Lieve Macken (2017). "Chapter 10: scate Taxonomy and Corpus of Machine Translation Errors". In: *Trends in E-Tools and Resources for Translators and Interpreters*. Leiden, The Netherlands: Brill — Rodopi. ISBN: 9789004351783.

Toledo Báez, Cristina, Moritz Schaeffer, and Michael Carl (2017). "Experiments in Non-Coherent Post-editing". In: *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*. Varna, Bulgaria: Association for Computational Linguistics, Shoumen, Bulgaria, pp. 11–20.

Toral, Antonio and Víctor M. Sánchez-Cartagena (2017). "A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pp. 1063–1073. URL: https://aclanthology.info/papers/E17-1100/e17-1100.

Van Brussel, Laura, Arda Tezcan, and Lieve Macken (2018). "A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch". eng. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari et al. Miyazaki, Japan: European Language Resources Association (ELRA), pp. 3799–3804. ISBN: 979-10-95546-00-9.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017a). "Attention is All you Need". In: *Proceedings of NIPS 2017*. Long Beach, CA, USA.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin (2017b). "Attention is All you Need".

In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vieira, Lucas N. (2016). *Cognitive Effort in Post-Editing of Machine Translation: Evidence from Eye Movements, Subjective Ratings, and Think-Aloud Protocols*. Newcastle University: PhD thesis.

Vilar, David, Jia Xu, Luis Fernando D'Haro, and Hermann Ney (2006). "Error Analysis of Statistical Machine Translation Output". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney (2016). "CharacTer: Translation Edit Rate on Character Level". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, pp. 505–510. DOI: 10.18653/v1/W16-2342. URL: https://www.aclweb.org/anthology/W16-2342.

Way, Andy (2018). "Quality expectations of machine translation". In: *CoRR* abs/1803.08409. arXiv: 1803.08409. URL: http://arxiv.org/abs/1803.08409.

Wu, Yonghui et al. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". In: *CoRR* abs/1609.08144. arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

Yamada, Masaru (2019). "The impact of Google Neural Machine Translation on Post-editing by student translators". In: *The Journal of Specialised Translation* (31), pp. 87–106.

Yang, X. Jessie, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah (2017). "Evaluating Effects of User Experience and System Transparency on Trust in Automation". In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '17. Vienna, Austria: ACM, pp. 408–416. ISBN: 978-1-4503-4336-7. DOI: 10.1145/2909824.3020230. URL: http://doi.acm.org/10.1145/2909824.3020230.

Yngve, Victor H. (1964). "Implications of Mechanical Translation Research". In: *Proceedings of the American Philosophical Society*. Vol. 108. 4. American Philosophical Society, pp. 275–281.

Zechner, Klaus and Alex Waibel (2000). "Minimizing Word Error Rate in Textual Summaries of Spoken Language". In: *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. URL: https://www.aclweb.org/anthology/A00-2025.

Appendices

Appendix A

Annotation guidelines

In this task, terms occurring in a data set of institutional academic texts, i.e. course unit descriptions and degree programmes, will be annotated. Annotations are carried out in the MT-EQuAl environment and the following fields are available:

- Target sentence (in English). This is the sentence where you will annotate terms.
- Source sentence (Italian).
- Information regarding the sentence domain (provided in the reference field).

To define what words are terms and what are not, the following definitions might help:
"A term is a graphic and/or phonic sign - a word or group of words, a compound word or a locution, an abbreviation - that allows to express a special concept related to concrete or abstract objects [...] that can be uniquely defined within a given discipline."[1].

Cabré and Sager (1999), maintained that "a terminological unit, or a term is a conventional symbol that represents a concept defined within a particular field of knowledge".

Also, notice that a term does NOT have to be a complex (group of) word(s). It can also be a simple/common (group of) word(s), if it is often used in a particular domain. For example, vehicle can be a term if is used in the domain of means of transport.

The text you are going to work on are NOT written by English native speakers. Therefore they might contain wrong terms. Please make sure that what you annotate as a term is also correct, e.g. by checking it in IATE, Google scholar or with a Google search.

Following these definitions of a term, please annotate terms belonging to the following **categories**:

- **No terms**: The sentence does not contain a term. In this case, highlight the whole sentence and select 'no term'.
- **Too many terms**: this is a default category in MT-EQuAl. Since we are interested in collecting as many terms as possible, try NOT to assign this category.
- **Sure term education (S-education)**: Use this label when you are sure that what you have found is a term and when the term belongs to the education domain, i.e. it is related to activities carried out inside an educational institution or to people being part of an educational institution.

[1] Quoted from: `https://bit.ly/2OpLClB`, my translation.

- **Possible term education (P-education)**: Use this label when you think that what you have found might be a term - but you are NOT completely sure about it - and when the term belongs to the education domain, i.e. it is related to activities carried out inside an educational institution or to people being part of an educational institution.

- **Sure term education equipment (S-education-equip)**: Use this label when you are sure that what you have found is a term and the term refers to educational equipment that could also be used elsewhere (e.g. overhead projector, desk, lab).

- **Possible term education equipment (P-education-equip)**: Use this label when you are NOT sure that what you have found is a term and the term refers to educational equipment that could also be used elsewhere (e.g. overhead projector, desk, lab).

- **Sure term disciplinary (S-disciplinary)**: Use this label when you are sure that what you have found is a term, and the term belongs to the discipline taught in that particular course. In the annotation interface, you will see information on the domain in the reference field, i.e. the name of the course unit and/or the name of the degree course, followed by the name of the general domain the course was assigned to.

- **Possible term disciplinary (P-disciplinary)**: Use this label when you are NOT sure that what you have found is a term, and the term belongs to the discipline taught in that particular course. In the annotation interface, you will see information on the domain in the reference field, i.e. the name of the course unit and/or the name of the degree course, followed by the name of the general domain the course was assigned to.

- **Wrong term**: If one term is wrong in the target text - e.g. if there is a wrong term/lexical choice or a typo - do not annotate it and write a comment to explain which term in the sentence you think is wrong.

Please pay attention to what follows:

- Many of the terms you will find are multi-word terms. When deciding the span of terms to be annotated, follow this principle: as short as possible, as many as necessary. Add words to a (multi-word) term only if they are needed to identify a concept.

- There is often a blurred line separating collocations from multi-word terms; for instance, consider carefully before annotating verbs as part of would-be multi-word terms (they are usually collocations!).

- When a whole term is followed by its acronym or abbreviation, annotate both the term and the short version separately.

- In some sentences you might find terms from different disciplinary domains - e.g. aerospace engineering sentences might include terms borrowed from algebra or mathematics. Annotate these as terms, independently from the domain in which they are used.

- For most sentences, the domain is provided in the reference field. However, be aware of the fact that for some sentences (more or less 20) no information on the domain is provided, because it was not possible to retrieve the whole text the sentence belonged to.

Appendix B

Experiment instructions

***NB**: The following instructions were provided to students assigned to the revision task. Identical instructions were provided to post-editors, except for the replacement of "Human translation" and "revision" with "Machine translation" and "post-editing".*

Human translation revision - Instructions
Please read this whole page carefully before starting with this task.

Introduction
In this task you will be asked to revise a single document containing two course unit descriptions translated from Italian into English by a human. Here are some instructions.

General guidelines:
The first course unit description you will work on describes a chemistry course: it starts with the source segment "CHIMICA". The second one describes a pharmacy course and starts with the source segment "FARMACIA".

- Work as you are used to! You are free to refer to any external resource (web sites, online dictionaries, etc.).
- Your productivity is very important for this experiment, so please:
  - Do not talk with other people during the task.
  - Ask questions only if strictly necessary.
  - While carrying out the task, try to focus as much as possible.
- Since your productivity will be measured on a segment level, please:
  - Always remember to confirm a segment and go to next one (Ctrl + Enter) as soon as you are done editing it.
  - Always make sure that the segment you are reading is activated, i.e. click on it and check that it is NOT greyed out.
- If a target segment reads "DO NOT TRANSLATE", this means that the segment is provided to make you understand the context of the following sentences, but you do NOT need to work on it.
- If during the task you see a "Translation conflict" signalled next to the Glossary tab, ignore it.

Instructions:

- Revise the text in order to make sure that the target text communicates the same meaning as the source text.
- At the end of the revision task, the text must be of a publishable quality for the website of a university. In particular:
  - Make sure that the use of key terminology is correct and consistent in the target text.
  - Make sure that grammar, syntax and spelling are correct.
- Do not edit based on stylistic preferences. If a segment is correct and publishable, leave it as it is, trying to use as much of the translated text as possible.
- To help you carry out this task, a manually validated termbase is added to the project. Please remember that:
  - Terms included in the termbase are underlined in red in the source sentence.
  - It is necessary to click on "Glossary" under the source sentence to be able to see the terms included in the termbase.

To start:

- Fill in the pre-experiment questionnaire.
- During the whole experiment, you will see a Microsoft Word window open. Ignore it and DO NOT close it.
- Open the file with the link to your project and click on the link/copy and paste it in Google Chrome.
- Now you can start working on your revision task (**IT** $>$ **EN-GB**) following the guidelines and instructions listed above.

When you are done:

- Make sure that you have confirmed all segments (the blue bar in the bottom-left part of MateCat editor window must be on 100%).
- Close the browser, but DO NOT turn off your computer.
- Fill in the post-experiment questionnaire and leave it on your desk.

Appendix C

Pre-experiment questionnaire

PRE-EXPERIMENT QUESTIONNAIRE ON MACHINE TRANSLATION AND REVISION

1) Please specify the email address used to access MateCat: mtexp._____@unibo.it
2) Please state your native language here: _____
3) How much experience do you have with **revision** outside the classroom?
   a) I have had no experience with revision outside the classroom.
   b) I have had little experience with revision outside the classroom (1-5 professional tasks).
   c) I have carried out a good amount of revision activities outside the classroom (more than 5 professional tasks).
4) How much experience do you have with **post-editing** outside the classroom?
   a) I have had no experience with post-editing outside the classroom.
   b) I have had little experience with post-editing outside the classroom (1-5 professional tasks).
   c) I have carried out a good amount of post-editing activities outside the classroom (more than 5 professional tasks).
5) How useful do you think Machine Translation is for translators?
   a) not useful
   b) useful
   c) very useful

Appendix D

Post-experiment questionnaire

POST-EXPERIMENT QUESTIONNAIRE ON MACHINE TRANSLATION AND REVISION

1) Please specify the email address used to access MateCat: mtexp._____@unibo.it
2) How would you rate the quality of the translation you revised?
   a) Poor: I had to edit almost every segment. Translating from scratch would have been faster.
   b) Sufficient: I had to edit most of the segments, but translating from scratch would have taken more time.
   c) Good: I only had to edit a few segments, and/or I only made minor changes.
3) How do you feel about the suggestions included in the termbase?
   a) I hardly ever looked at the termbase.
   b) I looked at the termbase suggestions when provided, but then double checked when I was not sure about the translation.
   c) I adopted the termbase suggestions when provided, without checking them further.
4) Only if your answer was *a* or *b*, please state the reason why you did not use/trust the termbase:
   a) I forgot that a termbase was available.
   b) I already knew the translation of the terms.
   c) The target text looked sufficiently reliable.
   d) Some termbase suggestions did not look reliable to me.
   e) Other: _____