MACHINE TRANSLATION AND CORPUS-BASED POST-EDITING IN THE FINANCIAL PRESS: A COMPARATIVE STUDY OF SMT AND NMT OUTPUTS

Patrizia Giampieri
University of Perugia, Italy

Abstract: *Machine translation (MT) has made huge strides in the last few decades. There are many strands of research dedicated to analysing the various types and advantages of MT, such as those related to statistical machine translation (SMT), and neural machine translation (NMT). Nonetheless, there are still a few shortcomings that the literature addresses and warns against. Also, little has been investigated regarding the quality of SMT or NMT with respect to corpus evidence. This paper aims to bridge this gap by exploring the quality and accuracy of SMT and NMT in the translation, from English into Italian, of a text in the financial field. To this aim, it compares MT outputs and corpus results by addressing an extract of a newspaper article sourced from the financial press. The paper findings report that NMT performs better in terms of word order and grammar accuracy, whereas SMT is more effective with naturally sounding words and sector-specific terms. The paper claims the high reliability of corpus data. As a matter of fact, not only does corpus evidence show language patterns and word usages in context, but it also suggests collocations and a variety of alternative terms in specific contexts.*

Keywords: *machine translation; corpus-based post-editing; statistical machine translation; neural machine translation; specialised translation; financial translation; corpus-assisted translation*

1. Machine translation

Machine translation (MT) is a process whereby computers translate texts automatically. There are currently two main MT approaches: statistical machine translation (SMT) (Brown et al.; Koehn et al.) and neural machine translation (NMT) (Kalchbrenner, Blunsom; Sutskever et al.; Bahdanau et al.).

1.1 Statistical machine translation

SMT considers natural language translation as a learning issue (Lopez 1), where MT delivers translations by analysing samples of authentic (i.e., human-made) translations (ibid.). SMT searches for correspondences between the source text and the target text from parallel corpora (Nießen, Ney 181). Each word of the target language is assigned a certain probability, and the highest probability is assumed to produce the best translation (Brown et al.).

Many scholars have contributed to the development of SMT to improve its quality and accuracy. For example, Nießen and Ney successfully propose incorporating morphological and syntactic information into SMT systems. Hardmeier compares approaches to discourse in SMT to identify existing issues. Banik et al. study a combination of statistical systems to exploit the advantages of existing MT tools.

1.2 Neural machine translation

Differently from SMT, NMT builds and uses a neural network of encoders and decoders, where the first ones encode a source sentence into a vector so that the latter can generate translations (Bahdanau et al.). In this way, the translation performance is maximized (ibid.). Thanks to this approach, NMT has reached cutting-edge results in several language pairs (Junczys-Dowmunt et al.). Researchers have investigated the potential and advantages of NMT. Amongst others, Riviera-Trigueros informs that NMT, in particular Google Translator, is the most used form of machine translation at the moment. With a view to improving MT performances, Haifeng et al. (2021) discuss how MT has made huge strides after the advent of NMT. They study state-of-the-art automatic translation methods (such as Speech-To-Text translation) and propose ways to further conceptualise MT tools. For example, Ghassemiazghandi et al. evaluate machine translation advancements on the basis of a Translation Error Rate (TER) and a Human-targeted Translation Error Rate (HTER).

1.3 SMT vs NMT

In light of the peculiarities outlined above, there are currently strands of research investigating whether and to what extent NMT and SMT differ and can produce different target texts. Koehn explains that NMT is more effective with large amounts of data and it performs better on low-frequency words (ibid.). In an effort to foreground differences among various MT systems, Cambedda et al. investigate the quality of SMT compared with NMT by analysing the automatic translation from Russian into Italian of medical articles using Yandex and DeepL. On the basis of their findings, they argue that DeepL (NMT) performs the overall translation better, but Yandex (SMT) is more accurate with culturally specific words. Diab posits that NMT produces fewer grammatical errors and mistranslations than SMT in the analysis of English-Arabic automatic translations, whereas Skadina and Pinnis claim that, in their English-Latvian post-editing study, NMT gives rise to a number of mistranslations and inaccuracies.

Despite the many advantages offered by machine-driven algorithms, Chan warns against some of the pitfalls of MT processes and highlights the importance of human intervention for quality purposes.

As can be seen, there are several approaches and findings, probably depending on the register or style of the source text, the field of interest, and the language pairs. Nonetheless, so far little research has been carried out as concerns corpus-assisted vs machine-based translations of articles sourced from financial newspapers. In other words, there is a paucity of systematic studies encompassing and comparing corpus-based and machine-driven translations in the financial field. This paper is aimed at filling this gap.

2. Corpora

Corpora are considered reliable language tools in translation training and practice (Zanettin; Bernardini; Bernardini, Ferraresi). They are collections of "authentic texts" (Bowker and Pearson 9), and, for this reason, they are considered particularly reliable, as they show patterns of real language. They contain samples of language usages in context (Farr, O'Keeffe) and help notice the nuances of second and/or sector-based language (Aston).

Scholars have carried out research to analyse the advantages of corpus-based translations, especially in specific fields of study, such as medicine, law and finance (Gavioli, Zanettin; Giampieri, Labruzzo Forshaw). Gavioli and Zanettin, for instance, report that corpora are helpful when addressing the phraseology of the medical field. The same is stated by other scholars, who postulate that corpora are particularly useful in technical and medical translator training because they raise users' awareness on sectoral language (see Sánchez Cárdenas, Faber). In the legal field, Giampieri posits that, to some extent, corpus consultation can compensate for the lack of specific knowledge. In the financial sector, the Hong Kong Financial Services Corpus (HKFSC) represents a milestone. It was developed by the Research Centre for Professional Communication in English of The Hong Kong Polytechnic University, and it is available in an annotated and non-annotated version. As far as corpus-based translations in economics and finance are concerned, Giampieri and Labruzzo Forshaw provide several hands-on examples with DIY (do-it-yourself) offline corpora (124-132), as well as with generic online corpus platforms (136-140). In particular, corpus-assisted translations prove to be satisfactory, and the language of target texts resembles the one of parallel (i.e., similar) texts in the source language.

In light of the above considerations, this paper is intended to shed light on the quality of the machine translation from English into Italian of a financial text sourced from an online financial newspaper. To this aim, a corpus of Italian news is consulted, and machine-generated output is compared with corpus evidence. At the same time, the possibility and accuracy of corpus-based post-editing is explored.

3. Research question

This paper is aimed at assessing the translation quality of NMT and SMT solutions by comparing the translation performances of the DeepL and Yandex platforms, respectively, with corpus evidence. The quality and reliability of corpus-driven post-editing is also investigated. The research questions that this paper wishes to address are the following: 1) Are NMT and/or SMT reliable in light of the authentic sectoral language represented in an *ad hoc* corpus? 2) To what extent can corpus consultation help tackle MT shortcomings and be used as a post-editing tool?

To answer these questions, a financial article is translated automatically from English into Italian by using the MT tools mentioned above. Automatically translated terms and phrases are then searched for and analysed in the CORIS (corpus of written Italian) (“press” sub-corpus) (Rossini Favretti). In this way, it is possible to assess the quality and performance of the two MT platforms as regards accuracy, fluency, authenticity, terminology, style and precision (Diab). At the same time, the possibility to post-edit MT output via corpus consultation is also investigated.

4. Methodology

This paper explores the quality of both the statistical and neural automatic translation (from English into Italian) of an extract of a newspaper article dealing with the 2016 Greek financial crisis. The source text is 150 words long and it was retrieved from the Bloomberg website (Chrysoloras, Gordon). Despite the reduced length of the article, it can be argued that the analysis of such a small text can yield linguistically rich insights because financial language is generally dense and highly patterned (Birzoim, 16).

The translation performances by the DeepL and Yandex platforms are assessed in light of corpus evidence. In particular, the “press” sub-section of the CORIS (corpus of written Italian) (in its annotated version) is consulted in order to verify whether the machine-translated terms and phrases are frequent and accurate. With regard to frequency, it should be noticed that occurrence is accounted for in the present study along with genre conventions and pragmatic appropriateness.

The CORIS is a freely accessible online corpus of Italian written documents. It is updated every three years and consists of a collection of authentic and frequently occurring texts selected on the basis of their representativeness of modern Italian. It is composed of 165 million words and is divided into several sub-sections, amongst which “press” is included. For the purpose of this paper, the “press” sub-corpus is selected.

5. Analysis

Table 1 below shows the source text. Idiomatic expressions and sector-based words or phrases are underlined.

Table 1. The source text

Greece Pressure Mounts as ECB Shows Caution on Bank Funds

Pressure mounted on Greece as U.S. and European officials called on the government to reach a deal with its creditors and the European Central Bank granted the nation’s cash-strapped banks only a small increase in emergency funds.

Greece has been at odds with other euro-area governments over the formula needed to extend the country’s 240 billion-euro rescue beyond its expiry at the end of February.

The country risks being left without a financial backstop and on course to default on some of its liabilities as early as next month if it doesn’t reach a creditor accord.

Documents outlining the government’s stance during two closed-door meetings with euro-area finance ministers and representatives of the so-called troika of the European Commission, the International Monetary Fund and the ECB, showed Athens is still seeking to radically alter the terms of the bailout memorandum.

As it can be noticed, there are plenty of idiomatic and fixed expressions, such as “pressure mounts/mounted (on)”, “shows caution on”, “called on”, “cash-strapped”, “been at odds”, and “closed-door meetings”. There are also technical or sector-based words, such as “the formula” (relating to a financial measure), “financial backstop”, “default”, “troika”, and “bailout memorandum”. It is now interesting to explore how MT tackles the above terms and phrases.

The next sections unveil MT shortcomings, if any, and address issues by means of corpus analysis. The tables that follow show the machine translations by DeepL and Yandex, respectively. The target text is divided into paragraphs to allow for better comprehension. Each table displays a paragraph. Potential translation issues are underlined and are investigated via corpus analysis. Table 2 below showcases the two automated translations of the title.

Table 2. The MT of the title

NMT (DeepL)	SMT (Yandex)
La pressione sulla Grecia aumenta mentre la BCE mostra cautela sui fondi bancari	*Grecia Pressione aumenta come BCE mostra cautela sui fondi bancari*

In Table 2 above, SMT proposes a literal rendering of the original text (e.g., “*Grecia Pressione*”, translating “Greece Pressure”). Written in this way, the noun phrase makes no sense in Italian. By writing *"pressione" ("sulla|sul")* in the CORIS search field, it is possible to explore whether “*pressione*” (back-translation: “pressure”) is followed by the preposition “*sulla*” or “*sul*” (back-translation of both: “on the”). As a result, concordances with “*pressione sulla Grecia*” are retrieved; this phrase is the same as the one proposed by NMT. Conversely, there is no corpus evidence of “*Grecia pressione*”.

The source expression “as ECB shows caution” is best rendered by NMT (i.e., “*mentre la BCE mostra cautela*”, back-translation: “while ECB shows caution”), in lieu of “*come BCE mostra cautela*” (back-translation:

“like ECB shows caution”), which is proposed by SMT. In the latter case, there are syntactical and cohesion issues. The SMT-driven target phrase, in fact, makes no sense in Italian. On the other hand, it is observable that there is no definite article preceding “ECB”, whereas NMT proposes “*la*”. By querying “*BCE*” in the CORIS, it becomes apparent that *BCE* is always preceded by the definite article “*la*”. Therefore, the target phrase by NMT is correct.

With regard to “as ECB shows caution on”, verbs preceding the Italian word “*cautela*” are searched for in the CORIS (search string: *[pos="V_GVRB"] "cautela"*). The verb “*mostrare*” (back-translation: “show”) comes to the fore, together with other verbs which are valid alternatives, such as “*esprimere*” (back-translation: “express”) and “*raccomandare*” (back-translation: “recommend”). Therefore, the machine-translated verb “*mostrare*” is confirmed by corpus evidence. As could be noticed, however, corpus analysis yields instances of various language patterns. Table 3 reports the MTs of the first paragraph.

Table 3. MTs of the first paragraph

NMT (DeepL)	SMT (Yandex)
La pressione è montata sulla Grecia, mentre i funzionari statunitensi ed europei hanno invitato il governo a raggiungere un accordo con i suoi creditori e la Banca centrale europea ha concesso alle banche della nazione in difficoltà solo un piccolo aumento dei fondi di emergenza.	*Pressione montata sulla Grecia come U. S. e funzionari europei ha invitato il governo a raggiungere un accordo con i suoi creditori e la Banca Centrale europea ha concesso banche a corto di liquidità della nazione solo un piccolo aumento dei fondi di emergenza.*

In Table 3, the verb “mounted” in the phrase “pressure mounted on Greece” is translated literally by both MT tools, which propose the verb “*montata*”. By querying the CORIS and by searching for any finite and non-finite verb collocating with the word “*pressione*” (search query: *"pressione" [pos="V_GVRB"]*), the following finites emerge: “*aumenta*” (back-translation: “rises”) and “*continua*” (back-translation: “continues”).

As there are no results with “*montata*”, it can be claimed that the automated target texts are inaccurate as the verb proposed is not frequent in authentic contexts. In addition, SMT features neither the use of the definite article (“*la*”) before “*pressione*”, nor the auxiliary verb “*è*” afterwards. As a matter of fact, the search string “*pressione aumentata*” generates no occurrences in the CORIS, whereas “*la pressione è aumentata*” does. Hence, in this case, NMT is more in line with corpus evidence.

The verb phrase "called on the government" is apparently addressed effectively by both MTs, which suggest "*invitato il governo*" (back-translation: "invited the government"). However, bilingual dictionaries[1] also frame "*fare appello a*" (back-translation: "appeal") as a translation option of "to call on". Therefore, both the verb "*invitare*" and the verb phrase "*fare appello*" can be queried in the corpus. By writing the following string in the CORIS: *("invitare|appello") []{0,2} "governo"*, it is possible to explore whether the verb "*invitare*" or the noun "*appello*" precede the word "*governo*" within a span of 0-2 words. Corpus results show that there are only two instances of "*invitare il governo*", whereas there are many phrases with the verbs "*fare*", "*rivolgere*" and "*lanciare*" preceding "*un appello al governo*". Therefore, although lexically correct, the two automatically generated target phrases do not accurately mirror the most frequent field-related expressions.

The phrase "the nation's cash-strapped banks" is translated as "*banche della nazione in difficoltà*" (back-translation: "national banks in difficulty") by NMT, whereas SMT suggests "*banche a corto di liquidità della nazione*" (back-translation: "cash-strapped banks of the nation"). Evidently, there are word order issues in the SMT-generated output. In particular, the position of the prepositional phrase "*della nazione*" is wrong as it should follow "*banche*". As regards corpus evidence, by searching for "*corto di*" followed by any common noun (search query: *"corto" "di" [pos="NN"]*), the following synonymous phrases appear: "*a corto di liquidi*"; "*a corto di liquido*"; "*a corto di contanti*", and "*a corto di liquidità*". The latter tends to outnumber the other options. Therefore, in this case, SMT produces a translation which mirrors authentic language, despite the incorrect word order. With regard to NMT, the phrase "*in difficoltà*" is freatured in the corpus. By querying the following string: *("banche"|"banca") []{0,5} "in" "difficoltà"*, equivalents of the word "bank" or "banks" are queried together with the Italian equivalent of "in difficulty" within a span of 0-5 words. Many hits with "*banche in difficoltà*" (back-translation: "banks in difficulty") are obtained. Nonetheless, it can be argued that the culturally bound expression "cash-strapped" is better rendered by a similar culturally bound phrase in Italian (i.e., "*a corto di liquidità*", as proposed by SMT).

The adjective "small" in the source phrase "a small increase in emergency funds" is translated as "*piccolo*" by both MTs. It is interesting to verify whether "*piccolo*" collocates with "*aumento di fondi*" (back-translation: "rise/increase in funds"). Therefore, the CORIS is consulted to retrieve modifiers preceding the word "*aumento*". By writing *[pos="ADJ"] "aumento"* in the corpus search field, the adjectives "*fittizio*" (back-translation:

[1] See for example the Hoepli online dictionary: https://dizionari.repubblica.it.

“fictitious/fake”) and “*lieve*” (back-translation: “light/soft”) are found, where “*lieve*” best renders “small”. In this case, neither of the MTs account for the best field-related collocations. Table 4 displays the MTs of the second paragraph.

Table 4. MTs of the second paragraph

NMT (DeepL)	SMT (Yandex)
La Grecia è stata in disaccordo con altri governi della zona euro sulla formula necessaria per estendere il salvataggio del paese da 240 miliardi di euro oltre la sua scadenza alla fine di febbraio.	*La Grecia è stata in contrasto con altri governi dell’area dell’euro sulla formula necessaria per estendere il salvataggio del paese da 240 miliardi di euro oltre la sua scadenza alla fine di febbraio.*

First of all, both MTs propose “*è stata*” as a literal translation of the present perfect form (“has been”). The correct tense to use in Italian is, by contrast, the present simple (i.e., “*è*”), as Greece was still “at odds” with other euro-area governments at the time the article was written and published. Also, the idiomatic expression “at odds” is rendered as “*in disaccordo*” by NMT and “*in contrasto*” by SMT. Both expressions are correct and naturally sounding in Italian. Nonetheless, it is worthwhile exploring whether the former is more frequent than the latter. By writing the following query in the CORIS search field: *"in" ("contrasto"|"disaccordo")*, it is possible to verify whether the preposition “*in*” is followed more frequently by “*disaccordo*” or “*contrasto*”. Corpus evidence shows that “*in contrasto*” greatly outnumbers “*in disaccordo*”. In this case, statistical machine translation performs better.

The expression “euro-area” is “*zona euro*” (back-translation: “euro zone”) in NMT and “*area dell’euro*” (back-translation: “area of the euro”) in SMT. By writing the following search string in the CORIS: *"euro" ("area"|"zona")*, the words “*area*” or “*zona*” are searched for after “*euro*”. The corpus mostly produces results with “*euro area*”. Conversely, there are no concordances with “*euro zona*”. Now it is sensible to verify whether “*euro*” follows “*area*” or “*zona*”. Therefore, the following syntax is written: *("area"|"zona") []{0,4} "euro"*. With this query, “*area*” or “*zona*” are investigated together with “*euro*” within a span of 0-4 words to the left. In this case, the expressions “*zona dell’euro*”, “*zona euro*”, “*area dell’euro*” and “*area euro*” are retrieved, where “*area dell’euro*” is the most frequent. This is another example where SMT is more accurate.

As can be seen in Table 4, both MTs propose “*formula necessaria*” as a translation of “formula needed”. If the string *"formula" "necessaria"* (back-translation: “formula necessary”) is searched for in the corpus, no hits are retrieved. By querying any adjective following “*formula*” (search string:

"formula" [pos="ADJ"]), no similar expressions to the source are obtained. The adjectives following “*formula*” are, in fact, “*migliore*”, “*ideale*”, “*magica*”, and “*operativa*” (back-translation: “better”, “ideal”, “magic” and “operative”). The corpus-sourced phrase “*formula migliore*” (back-translation: “better formula”), although apparently satisfactory, does not relate to financial matters (e.g., “*la formula migliore per gestire il sonno*”, back-translation: “the best way/formula to deal with sleepiness”). It could be assumed that the calque “*formula*” might not be the right translation option in context. Therefore, a possible equivalent of “needed” (in the expression “formula needed”), such as “*necessario/a*”, is investigated. By writing *[pos="NN"] ("necessario"|"necessaria")* in the CORIS search field, nouns preceding “*necessario/a*” are obtained, such as “*condizione*”, “*manovra*” and “*misura*” (back-translation: “condition”, “manoeuvre” and “measure”, respectively). In light of these results, an acceptable translation of “formula needed” is “*manovra necessaria*”.

The verb “*estendere*” is proposed by both MT platforms to render “to extend”. However, *estendere* could be a debatable translation option as it may not relate to debts or bailouts, and it could not collocate with “*scadenza*” (“expiry”). By searching for any verb preceding “*scadenza*” in the CORIS (search query: *[pos="V_GVRB"] []{0,6} "scadenza"*), the following phrases are retrieved: “*prosegua dopo la scadenza*” (back-translation: “continue after the deadline/expiry”); “*proseguire oltre i termini della scadenza*” (back-translation: “continue beyond the expiry/deadline’s terms”); “*prorogare alla scadenza*” (back-translation: “extend after the expiry”), and “*si spera in una possibile proroga alla scadenza*” (back-translation: “we hope on a possible extension at the expiry”). Therefore, in light of corpus evidence, possible translations of “(to) extend beyond its expiry” are “*proseguire dopo la scadenza*” and “*prorogare alla scadenza*”. Differently from MTs, no possessive adjective is used before “*scadenza*”.

The translations of the third paragraph are shown in Table 5 below. Sector-based phrases “financial backstop” and “on course to default” may be challenging to translate.

Table 5. MTs of the third paragraph

NMT (DeepL)	SMT (Yandex)
Il paese rischia di rimanere senza un supporto finanziario e di andare in default su alcune delle sue passività già il mese prossimo, se non raggiunge un accordo con i creditori.	*Il paese rischia di essere lasciato senza un backstop finanziario e in rotta di default su alcune delle sue passività già il mese prossimo se non raggiunge un accordo con i creditori.*

As can be seen, NMT proposes “*supporto finanziario e di andare in default*”, whereas SMT is more literal and suggests “*backstop finanziario e in rotta di default*”. Bilingual dictionaries suggest “*protezione*” (back-translation: “protection”) as a translation of “backstop”. If the following search string is queried in the CORIS, *("supporto|protezione|backstop") ("finanziario|finanziaria")*, it is possible to notice that “*supporto*” outnumbers “*protezione*”, whereas there are no occurrences of “backstop”. Therefore, the term proposed by NMT seems more accurate. In addition, by searching for a synonym of “*supporto*”; i.e., “*sostegno*” (search string: *"supporto|sostegno" "finanziario"*), the word “*sostegno*” greatly outnumbers “*supporto*”. The best translation of “financial backstop” is, hence, the phrase “*sostegno finanziario*”.

In the expression “on course to default”, the prepositional phrase “on course to” can be rendered with the idiomatic “*in rotta di*”, as proposed by SMT. However, as a precaution, it is useful to verify whether “*in rotta di*” collocates with “default” (which is an English borrowing used in Italian financial texts). If the word “default” is analysed in the corpus, other collocations emerge, such as “*verso il default*” in the verb phrase “*precipitando verso il default*” (back-translation: “falling/rushing into default”), or “*andare in default*”, which is also suggested by NMT. The expression “*in rotta di default*” is not featured in the corpus.

In the source text, the word “default” is followed by “liabilities”, which is “*passività*” in both MT target texts. Dictionaries may propose other translations of “liabilities”, such as “*debiti*” (back-translation: “debts”). If the string *"default" []{0,10} ("passività|debiti|debito")* is written in the CORIS search field, several occurrences with the following phrase are obtained: “*default sul debito*” (back-translation: “default on the debt”). There is only one hit with “*passività*”, which, however, is unrelated (i.e., “*default del gruppo che conta passività per 40 mila miliardi*”; back-translation: “the group’s default whose liabilities amount to 40 thousand billion”).

As regards the noun phrase “creditor accord”, both automatic translations are acceptable (i.e., “*accordo con i creditori*”). The corpus, in fact, generates concordances with this expression. However, corpus-sourced alternative solutions might be explored. By querying the phrase *"con" "i" "creditori"*, it is possible to noticed that it is preceded not only by “*accordo*”, but also by “*concordato*” (back-translation: “agreement”); “*negoziato*” (back-translation: “negotiation”), and “*compromesso*” (back-translation: “compromise”), which are all acceptable equivalents. Table 6 reports the MTs of the fourth and last paragraph.

Table 6. MTs of the fourth and last paragraph

NMT (DeepL)	SMT (Yandex)

I documenti che delineano la posizione del governo durante due riunioni a porte chiuse con i ministri delle finanze della zona euro e i rappresentanti della cosiddetta troika della Commissione europea, il Fondo monetario internazionale e la BCE, hanno mostrato che Atene sta ancora cercando di modificare radicalmente i termini del memorandum di salvataggio.	*I documenti che delineano la posizione del governo durante due riunioni a porte chiuse con i ministri delle finanze dell'area dell'euro e i rappresentanti della cosiddetta troika della Commissione europea, del Fondo monetario internazionale e della BCE, hanno mostrato che Atene sta ancora cercando di modificare radicalmente i termini del memorandum di salvataggio.*

In the last paragraph, the phrase “the documents outlining the government’s stance” is rendered as “*i documenti che delineanon la posizione del governo*” by both MT tools. It would be wise to explore whether the verb “*delineano*” collocates with “*documenti*”. To do so, the phrase *"i" "documenti" "che" [pos="V_GVRB"]* is searched for. Interestingly, there are no instances of the verb “*delineano*”. Conversely, many hits show the following verbs: “*contengono*” (back-translation: “contain”); “*riflettono*” (back-translation: “reflect”); “*concludono*” (back-translation: “conclude”) (which is, however, unrelated); “*indicano*” (back-translation: “indicate”), and “*definiscono*” (back-translation: “define”). Also, if the search string *[pos="V_GVRB"] "la" "posizione" "del" "governo"* is queried, the verbs preceding the phrase “*la posizione del governo*” are generated, and the following ones come to the fore: “*definire*”, “*chiarire*”, “*riassumere*” (back-translations: “define”, “clarify” and “summarise”). This is another example where corpus analysis provides a variety of translation options.

As regards the idiomatic expression “door-closed meeting”, both MTs are accurate as corpus evidence confirms the automated translations proposed (i.e., “*riunione a porte chiuse*”).

Concerning the phrase “representatives of the so-called troika”, both MT outputs contain “*rappresentanti della cosiddetta troika*”. By searching for *"della" []{0,2} "troika"*, the following phrases are found: “*ambasciatori della troika*” and “*rappresentanti in Grecia della cosiddetta 'troika'*”. Therefore, it can be stated that both “*ambasciatori*” and “*rappresentanti*” are acceptable equivalents.

The verb “showed” refers to “documents” and is rendered literally (i.e., “*hanno mostrato*”) by both MT tools. However, by listing the verbs following the word “*documenti*” (search string: *"i" "documenti" [pos="V_GVRB"]*), the verbs obtained are as follows: “*rivelano*” (back-translation: “reveal”); “*precisano*” (back-translation: “specify”); “*dimostrano*” (back-translation:

"demonstrate"), together with many others. There is no occurrence of "*mostrano*".

The last phrase to analyse is "radically alter the terms of the bailout memorandum". The first part of the phrase is rendered as "*modificare radicalmente*" by both MT tools. Synonyms of "*modificare*" (e.g., "*alterare*" and "*cambiare*") may be queried. By searching for *("alterare"|"modificare"|"cambiare") "radicalmente"* in the CORIS, there are many hits where "*cambiare*" collocates with "*radicalmente*". As concerns the second part of the phrase, i.e., "bailout memorandum", both MTs propose "*memorandum di salvataggio*". Either "memorandum" or "bailout" can be transferred (i.e., used) in Italian, as they are frequent borrowings meaning "*accordo*" and "*salvataggio*", respectively[2]. Therefore, the query *("memorandum|accordo") []{0,5} ("salvataggio|bailout")* are written in the CORIS search field. The results, however, only show "*accordo sul/di salvataggio*" (back-translation: "memorandum/agreement on/of bailout").

6. Discussion

The analysis carried out in the sections above assessed the quality of the two MT outputs as regards accuracy, authenticity, fluency, terminology, style and precision (Diab). In light of the results obtained, it is evident that both the DeepL and Yandex tools are accurate (especially NMT, with regard to grammar and word order), but also imprecise. Sometimes both MTs neglect recurrent collocates and/or the words most frequently used in a particular context. Therefore, some improvements or fine-tuning are necessary.

In order to better focus on the text accuracy, preciseness, fluency, terminology and style, Table 7, 8 and 9 highlight the differences in the proposed translations by dividing mistranslations (or inaccurate translations) into three categories: a) wrong translations due to grammatical issues, word order and/or wrong lexical choices; b) terms proposed by MTs which are less frequent in the Italian language as a whole (as evidenced by corpus consultation), and c) MT inaccuracies due to less frequent sector-based terms (also assessed on the basis of corpus evidence). Table 7, 8 and 9 below report these findings (mistranslations are underlined).

As mentioned, Table 7 addresses the grammatical issues of both MTs, Table 8 reports less frequent terms proposed by MTs, and Table 9 shows MT shortcomings due to infrequent sector-based terms. Some translation infelicity may overlap and belong to more categories.

[2] By querying, for example, *bailout memorandum site: ilsole24ore.com* on Google, it is possible to retrieve webpages of the *IlSole24Ore* Italian financial journal where both *bailout* and *memorandum* are used.

Table 7. Source text, corpus-based translations and grammar issues in the SMT and NMT (the issues are underlined)

Source Text	**Corpus-based Translation**	**NMT (DeepL)**	**SMT (Yandex)**
Greece Pressure	*Pressione sulla Grecia*	*Pressione sulla Grecia*	*Grecia pressione*
As ECB	*Non appena la BCE*	*Mentre la BCE*	*Come BCE*
Mounted	*Aumenta*	*Montata*	*Montata*
Nation’s (…) banks	*Banche della nazione*	*Banche della nazione*	*Banche... della nazione*
Has been	*È*	*È stata*	*È stata*
Backstop	*Sostegno / Supporto*	*Supporto*	*Backstop*
On course to (default)	*Verso (il default)*	*In (default)*	*In rotta di (default)*

As evident in Table 7, statistical machine translation generates more grammatical inaccuracies. Word order is sometimes wrong (as in “*banche… della nazione*”), and literal renderings hinder comprehension (e.g., “*Grecia Pressione*” and “*come BCE*”), or sound unnatural (e.g., *montata*, “mounted”, referring to *pressione*, “pressure”). As a whole, neural machine translation produces more instances of error-free language patterns and seems in line with the (authentic) language of the corpus.

Table 8. Source text, corpus-based translations and less frequent terms (in the Italian language as a whole) proposed by SMT and NMT (the issues are underlined)

Source Text	**Corpus-based Translation**	**NMT (DeepL)**	**SMT (Yandex)**
Called on (the government)	*Fatto appello / rivolto / lanciato*	*Invitato*	*Invitato*
At odds	*In contrasto*	*In disaccordo*	*In contrasto*
Euro area	*Area euro*	*Zona Euro*	*Area Euro*
Its (expiry)	*- (scadenza)*	*Sua (scadenza)*	*Sua (scadenza)*

As can be seen from Table 8, neural machine translation tends to propose terms which are less frequent in the Italian language.

For example, differently from English, Italian does not need possessive adjectives before “*scadenza*”, as proved by corpus evidence (see the last line in Table 8 above).

Table 9. Source text, corpus-based translations and MT of sector-based terms (the issues are underlined)

Source Text	**Corpus-based translation**	**NMT (DeepL)**	**SMT (Yandex)**
Called on (the government)	*Fatto appello / rivolto / lanciato*	*Invitato*	*Invitato*
Cash-strapped	*A corto di liquidità*	*In difficoltà*	*A corto di liquidità*
Accord	*Accordo / Negoziato / Concordato / Compromesso*	*Accordo*	*Accordo*
Small (increase)	*Lieve (aumento)*	*Piccolo (aumento)*	*Piccolo (aumento)*
Formula	*Manovra*	*Formula*	*Formula*
Extend (rescue)	*Prorogare*	*Estendere*	*Estendere*
Liabilities	*Debiti*	*Passività*	*Passività*
(Documents) outlining	*Riflettono / Contengono / Indicano / Definiscono / Chiariscono*	*Delineano*	*Delineano*
(Documents) showed	*Rivelano / Precisano / Dimostrano*	*Mostrato*	*Mostrato*
Alter	*Cambiare*	*Modificare*	*Modificare*
Bailout memorandum	*Accordo (di Salvataggio)*	*Memorandum (di salvataggio)*	*Memorandum (di salvataggio)*

As far as Table 9 is concerned, both automatic translations render terms literally, without considering the context where they are employed. In addition, they neglect potential collocations. For example, the word “liabilities” is best rendered as “*debiti*”, instead of “*passività*”, especially in the sense given in the article. The same can be said of “*estendere*”, which does not mean “extend” in the given context and does not collocate with “*scadenza*” (“expiry”).

However, statistical machine translation performs more satisfactorily with regard to idiomatic expressions. For example, the phrase “cash-strapped” is better rendered as “*a corto di liquidità*” (SMT) instead of “*in difficoltà*” (NMT). The expression “*in difficoltà*” is not a mistranslation or a false equivalence, but the SMT target term is more precise, as it proposes an equivalent idiomatic expression.

The words and phrases suggested by MT in Table 9 are not wrong or inaccurate *per se*. There are neither mistranslations nor awkward constructions. However, given the contexts and the sector-based language, the terms need revision.

Finally, as can be noticed in Table 9, not only can corpus evidence help tackle MT shortcomings, but it also provides several translation options (see, for example, the translations of “backstop” in Table 7, or the several renderings of “called on”, “accord”, “outlining”, and “showed” in Table 9).

7. Conclusions

The paper was aimed at exploring whether the NMT and SMT of a financial text could be accurate and reliable, and whether it would resemble authentic texts. To do so, an English text dealing with the Greek financial crisis of 2016 was translated automatically into Italian by using the DeepL and Yandex tools. The two automated translations were then analysed in depth by consulting the CORIS (corpus of written Italian), which was used as a post-editing tool.

The paper findings show that NMT performed better as concerns grammar and word order. On the other hand, SMT mirrored naturally occurring language, but only on one occasion (i.e., in the rendering of the idiomatic “cash-strapped banks”). Both MT tools produced inaccuracies, mistranslations and infrequent renderings or collocations.

Finally, it should be noticed that corpus-assisted post-editing was successfully performed. Corpus evidence addressed MT shortcomings effectively, and it also provided many samples of naturally occurring and sector-related language. In addition, it produced a variety of alternative translation options in context.

The implications of this study in translator training are manifold. The analyses offered empirical evidence of how corpus consultation can provide insights into specialised discourse. This paper showcased how translator trainees can search for words in a sector-oriented corpus to retrieve occurrences, language patterns, and, hence, make informed decisions. To some extent, the investigation carried out also highlighted the trade-off between linguistic forms and field-related knowledge, thereby underscoring the need for ad hoc language tools that help novice or inexperienced translators cope with the difficulties of specialised language. From a pedagogical perspective, this paper also showed how corpus users can recognise and address recurrent machine translation errors, thus leveraging corpora as post-editing tools. In this way, automated translation is not perceived as a replacement for human intelligence or knowledge, but as a resource that requires expertise, critical assessment, and supervision.

The limits of this paper lie in the fact that only one financial article was focused on and only two MT platforms were taken into account. Future research could explore whether similar performances are obtained by different MT tools addressing a larger number of financial articles. Researchers could also examine whether similar evidence is produced when considering other sectors. The findings obtained in this paper, in fact, are initial and tentative. Further research is called on in order to corroborate or challenge them. Additionally, scholars may consider the future inclusion of human evaluation, more texts, and/or corpus triangulation analysed.

Works Cited

Aston, Guy. *Learning with Corpora*. Houston, TX: Athelstan, 2001.

Bahdanau, Dzmitry, Kyunghyun Cho, Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” *Proceedings of ICLR*, 2015. doi.org/10.48550/arXiv.1409.0473.

Banik, Denajyoty, Asif Ekbal, Pushpak Bhattacharyya, Siddhartha Bhattacharyya, Jan Platos. “Statistical-based system combination approach to gain advantages over different machine translation systems.” *Heliyon* 5(9) (2019). https://doi.org/10.1016/j.heliyon.2019.e02504.

Bernardini, Silvia. “How to use corpora for translation.” *The Routledge Handbook of Corpus Linguistics. Second edition.* Ed. Anne O’Keeffe, Michael J. McCarthy. Abingdon: Routledge, 2022. 485-498.

Bernardini, Silvia, Adriano Ferraresi. “Corpus linguistics”. *The Routledge Handbook of Translation and Methodology*. Ed. Federico Zanettin, Christopher Rundle. Abingdon: Routledge, 2022. 207-222.

Birzoim, Ammoon. “Cross-Domain Applications of LLM-Based Retrieval and Dialogue Systems: A Review of Current Practice.” *TechRxiv* (2025). https://doi.org/10.36227/techrxiv.175756155.57191866/v1.

Bowker, Lynne, Jennifer Pearson. *Working with specialized language: a practical guide to using corpora*. London/New York: Routledge, 2002.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Deppa Pietra, Frederick Jelinek, John D. Lafferty, Robert Mercer, Paul S. Roossin. “A statistical approach to machine translation.” *Computational linguistics* 16(2) (1990): 79–85.

Cambedda, Giulia, Giorgio M. Di Nunzio, Viviana Nosilia. “A Study on Automatic Machine Translation Tools: A Comparative Error Analysis Between DeepL and Yandex for Russian-Italian Medical Translation.” *Umanistica Digitale* 10 (2021): 139–163. https://doi.org/10.6092/issn.2532-8816/12631.

Chan, Sin-Wai. *The Human Factor in Machine Translation*. London/New York: Routledge, 2018.

Chrysoloras, Nikos, Paul Gordon. *Greece Pressure Mounts as ECB Shows Caution on Bank Funds*, 2015. https://www.bloomberg.com/news/articles/2015-02-18/greece-pressure-mounts-as-ecb-shows-caution-on-bank-funds.

Diab, Nessma. “Out of the BLEU: An Error Analysis of Statistical and Neural Machine Translation of WikiHow Articles from English into Arabic.” *CDELT Occasional Papers in the Development of English Education* 75(1) (2021): 181-211. https://doi.org/10.21608/opde.2021.208437.

Farr, Fiona, Anne O’Keeffe. “Using corpora to analyse language.” In *Routledge Handbook of English Language Teacher Education.* Ed.

Steve Walsh and Steve Mann. London/New York: Routledge, 2019. 268-282.

Gavioli, Laura, Federico Zanettin. “I corpora bilingui nell’apprendimento della traduzione. Riflessioni su un’esperienza pedagogica.” *I corpora nella didattica della traduzione* (*Corpus Use and Learning to Translate*). Ed. Silvia Bernardini, Federico Zanettin. Bologna: Cooperativa Libraria Universitaria Editrice Bologna, 2000. 31-44.

Ghassemiazghandi, Mozhgan, Tengku Sepora, Tengku Mahadi. “Quality estimation of machine translation for literature.” *The Human Factor in Machine Translation.* Ed. Sin-Wai Chan. London/New York: Routledge, 2018. 183-208.

Giampieri, Patrizia. “Can corpus consultation compensate for lack of knowledge in legal translation training?” *Comparative Legilinguistics* 46 (2021): 5-35.

Giampieri, Patrizia, Claudia Labruzzo Forshaw. *Technical Translations: a corpus approach for Italian and English speakers*. Turin: Celid, 2021.

Haifeng, Wang, Wu Hua, He Zhongjun, Huang Liang, Kenneth Church Ward. “Progress in Machine Translation.” *Engineering* (2021). https://doi.org/10.1016/j.eng.2021.03.023

Hardmeier, Christian. “Discourse in Statistical Machine Translation.” *Discours* 11 (2012). https://doi.org/10.4000/discours.8726.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, Hieu Hoang. “Is Neural Machine Translation Ready for Deployment? a Case Study on 30 Translation Directions.” *ArXiv* (2016). https://doi.org/10.48550/arXiv.1610.01108.

Kalchbrenner, Nal, Phil Blunsom. “Recurrent continuous translation models.” *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013): 1700–1709.

Koehn, Philipp. *Neural Machine Translation*. Cambridge: Cambridge University Press, 2020.

Koehn, Philipp, Franz J. Och, Daniel, Marcu. “Statistical Phrase-Based Translation.” *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (2003): 127-133.

Lopez, Adam. Statistical Machine Translation. *ACM Computing Surveys* 40(3) (2008): 1-49 https://doi.org/10.1145/1380584.1380586.

Nießen, Sonja, Hermann Ney. “Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information.” *Computational Linguistics* 30(2) (2004): 181–204. https://doi.org/10.1162/089120104323093285.

Rivera-Trigueros, Irene. "Machine translation systems and quality assessment: a systematic review." *Language Resources & Evaluation* (2021). https://doi.org/10.1007/s10579-021-09537-5.

Rossini Favretti, Rema. (2000). "Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS." *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Ed. Rema Rossini Favretti. Bulzoni, Roma, 2000. 39-56. https://corpora.ficlit.unibo.it/CORISPubs/Rossini2000_ProgettazioneeCostruzione.pdf.

Sánchez Cárdenas, Beatriz, Pamela Faber. "Corpus Analysis and the Translation of Adverbs in Specialised Texts: Raising Student Awareness." *Corpus-Based Approaches to Translation and Interpreting. From Theory to Applications.* Ed. Gloria Corpas Pastor, Míriam Seghiri. Frankfurt: Peter Lang, 2016. 195-217.

Skadiņa Inguna, Mārcis Pinnis. "NMT or SMT: Case Study of a Narrow-domain English-Latvian Post-editing Project." *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (2017): 373–383. Taipei, Taiwan.

Sutskever, Ilya, Oril Vinyals, Quoc V. Le. "Sequence to sequence learning with neural networks." *Proceedings of NeurIPS*, (2014): 3104–3112.

Zanettin, Federico. 2014. "Corpora in Translation." *Translation: A Multidisciplinary Approach.* Ed. Juliane House. London: Palgrave Macmillan, 2014. 178-199.

Online resources

CORIS: https://corpora.ficlit.unibo.it/TCORIS/.

DeepL: https://www.deepl.com/translator.

Hoepli online dictionary: https://dizionari.repubblica.it.

Hong Kong Financial Service Corpus http://rcpce.engl.polyu.edu.hk/HKFSC/?menu=..%2FHKFSC%2F.

Il Sole 24 Ore: https://www.ilsole24ore.com//.

Yandex: https://translate.yandex.com/.