Annali del Dipartimento di Metodi e Modelli per l'Economia il Territorio e la Finanza
2025, DOI: 10.13133/2611-6634/1749

E-ISSN 2611-6634

ISSN 2385-0825

Sapienza
Università Editrice

Research paper

First published online: October 27, 2025

Patrizia Giampieri[*]

AUTOMATIC TRANSLATION AND CORPUS ANALYSIS IN THE GENERATION OF ACADEMIC ABSTRACTS

Abstract

Machine translation (MT) has made huge strides in the last decades and it is increasingly applied in the ESL classroom, both for language learning and translation purposes. This paper wishes to explore whether and to what extent automatic translations performed by an MT platform and a Language Model (LM) tool can be effectively integrated and/or post-edited by corpus evidence. The language pair considered for the analysis is Italian/English, and the source text is an abstract focusing on academic Italian as an L2. The corpus consulted is the ARC (Anthology Reference Corpus), available on Sketch Engine. In this case, the Italian abstract is accompanied by an official translation into English. Therefore, this paper compares MT- and LM-driven output with the official target text. Additionally, it investigates to what extent corpus consultation can be integrated into the post-editing process to produce a qualitatively acceptable text in the target language. The paper's findings indicate the high reliability of corpus-driven post-editing, revealing alternative translation options. It also brings collocations to the fore, thereby foregrounding language patterns. In addition, corpus analysis helps address MT- and LM-driven shortcomings. The paper discusses how corpus-driven post-editing can be seamlessly incorporated into the translation process and into translation and language education.

Keywords: corpus analysis; corpus-based translation; machine translation; translation of academic texts

[*] Universitas Mercatorum, Italy.

1 MT in the context of EAP

English for Academic Purposes (EAP) is a branch of English for Special Purposes (ESP) that investigates and responds to the communicative needs of academic scholars, lecturers and students (Hyland and Hamp-Lyons 2002: 2). Academic language is studied, taught and focused on in both the written (see Hyland 2015) and spoken form (see Fortanet-Gómez 2006; Molino 2014). For the purposes of this paper, written academic English is addressed. Written EAP is characterised by specific features, or writing strategies, such as hedging (Hyland 1995), stance and engagement (Hyland 2011), as well as reformulations and exemplifications (Hyland 2007). Through hedging, for example, tentativeness and probability are formulated. In such manner, hedged phrases help writers express what they think in a polite manner, i.e., by using lessening words which render statements more cautious (Giampieri 2017: 54-55). By using engagement devices, writers aim to establish a connection with the readership and anticipate possible oppositions. In this way, they are able to create solidarity (Giampieri 2017: 56). Through stance, writers express their own opinions and attitude (ibid.: 57). Reformulations and exemplifications are strategies which help readers understand what is expressed in a text or passage (ibid.: 59). Given such peculiarities, EAP cannot be taught and learnt by simply "fixing up grammar" (Hyland and Hamp-Lyons 2002: 6). At the same time, the comprehension and use of field-related terminology should also be focused on (Cohen et al. 1988). Teaching EAP literacy is often carried out by means of corpora, and a great deal of corpus-based materials have been developed in the last decades (Mauranen 2003; Fortanet-Gómez 2006; Swales 2006; Molino 2014; Flowerdew 2015; Hyland 2015). Corpora are generally composed of authentic texts (that is, texts that are not necessarily intended for L2 learners). Therefore, they can provide examples of real (i.e., natural-occurring) language (O'Keeffe et al. 2007: 26). As a matter of fact, corpora allow users to notice collocations (i.e., words that frequently co-occur, Sinclair 1991), recurrent language patterns, as well as word usages in context. For these reasons, corpus-based EAP courses are usually designed for both native and non-native speakers of English (NS and NNS, respectively) (Morton 1999; Dunleavy 2003; Mauranen 2003 and 2012; Crème and Lea 2008; Randaccio 2013). There are several works dedicated to improving academic grammar knowledge and writing skills via corpora (Lee and Swales 2006; Biber and Conrad 2009). Some scholars have also planned EAP lessons and courses which envisage the assignment and/or performance of translation tasks (Shei 2005; Siegel 2023). In this way, they posit, both translation and writing competences are fostered (Shei 2005). When addressing intended usages of translations, for example, Scarpa (2020) defines abstracts as derived documents who condense an original academic work by offering an overview of the content. Additionally, their purpose may shift from the original text to embrace more accessible usages. For example, they could be translated for popular-science journals, thus entailing a change in genre. López-Arroyo and Méndez-Cendón (2007) scrutinise abstracts through the lenses of discourse conventions. They compare rhetorical and phraseological strategies in medical abstracts addressing diagnostic imaging in the English/Spanish language pair. Their findings indicate that Spanish abstracts are generally less formulaic and more varied than their English counterparts, thus revealing interesting and different language-driven approaches. Götz (2015) focuses on the field of Translation Studies and analyses the rhetorical strategies developed in original and translated abstracts. More precisely, she investigates abstracts translated from Hungarian into English and compares them with the original abstracts written in English. The author finds no relevant rhetorical difference, although original English abstracts feature more emphatic introductory opening, probably due to the importance of positioning research within the English international context. Strains of research have been dedicated to the potentialities of teaching post-editing techniques in machine translation (MT) assignments by using MT for second language (L2) skills improvement. Lee (2023), for example, provides useful insights in the application of MT in the L2 training of language lecturers. O'Brien et al. (2018) base their study on the assumption that NNS generally find it difficult to publish academic papers in an L2. For this reason, the authors investigate the advantages of the machine translation and post-editing of academic abstracts. In their findings, they claim that MT combined with post-editing activities produce qualitative outputs. In addition, the students' cognitive burden is fairly reduced. Lee (2019) focuses research studies on exploring whether MT can improve students' writing skills. To do so, academic students are prompted to translate their L1 writing into L2 without the assistance of MT and then to correct their L2 writing by using MT output for comparison. The findings show the usefulness of MT in helping students gain and/or improve academic literacy. However, the author argues that the lecturer's guidance is necessary to allow learners focus on language content and form (Lee 2019). Olkhovska and Frolova (2020) undertake an observation study with undergraduate students in Foreign Languages. They divide the students into two groups: the first one translates a text by using an MT tool, whereas the other group translates the same text without any CAT tools. The results of their findings confirm that without prior training in post-editing, students are not able to analyse translation options critically. Hence, their translation output tends to be poorer than the one of the students who translate the same source text manually. Given the huge strides forward made by MT and LM software applications, this paper is aimed at exploring the usefulness and possible applications of MT and LM (complemented by post-editing) in EAP. To this aim, it envisages the automatic translation (from Italian into English) of an academic abstract dealing with teaching/learning Italian as an L2. The quality assessment of the MT/LM output is carried out by comparing the automatically generated target texts with the official translation into English, as well as with corpus-based evidence. In this respect, the Association for Computational Linguistics (ACL) Anthology Reference Corpus (ARC) available on Sketch Engine (SE) (Kilgarrif et al. 2004) is consulted. The ARC is an English corpus of conference and journal papers dealing with Natural Language Processing and computational linguistics. The corpus is composed of over 62 million words.

2 Research questions

The research questions this paper wishes to address are the following:

1. How can the automatic translation of an academic abstract be used in translator training to foster both writing and translation skills?

2. How can corpus-based post-editing activities of an automatically translated text be integrated in the EAP classroom?

3. Does the corpus produce satisfactory evidence to be successfully integrated in the automated translation process?

To answer these questions, the English MT- and LM-based translations of an Italian abstract are compared with the abstract official English translation and with corpus evidence. For the purposes of this paper, as already mentioned, the Anthology Reference Corpus hosted on Sketch Engine is consulted.

3 Methodology

Table 1 below reports the abstract addressed. The words or phrases in bold are investigated via corpus analysis.

Table 1. Abstract.

> *NELLA* ***CLASSE A MICROLINGUE DIFFERENZIATE. PROGETTAZIONE DIDATTICA*** *PER L'ITALIANO LINGUA SECONDA, DI* ***STUDIO E DI SPECIALIZZAZIONE DISCIPLINARE***
>
> *Questo articolo* ***argomenta*** *la possibilità di* ***contemperare*** *l'apprendimento dell'****italiano generale*** *con quello dell'italiano come lingua dello studio e di* ***microlingue*** *specialistiche nella formazione di studentesse e* ***studenti universitari di madrelingua non italiana*** *in corsi* ***di livello A2 e B1*** *del Quadro comune europeo di riferimento per la conoscenza delle lingue.*
>
> ***Dati un monte ore*** *e una classe (quest'ultima con* ***determinati livelli di conoscenza in ingresso, provenienza geografica spesso mista, afferenza dei partecipanti a diversi corsi di laurea****),* ***come includere in un corso almeno un primo approccio alle microlingue scientifico-disciplinari di interesse delle studentesse/studenti?***
>
> *Il lavoro di progettazione può rispondere a questa domanda concentrando la propria azione sul design del curricolo e di alcune* ***attività didattiche*** *specifiche. Nel* ***delineare questa possibilità****, questo articolo intende* ***fornire uno spunto utile*** *a quelle/i insegnanti e coordinatrici/ori che* ***si trovano a operare entro precisi vincoli organizzativi****, senza voler rinunciare a un'offerta didattica che tenga conto delle plurime dimensioni in cui studentesse e studenti di madrelingua non italiana hanno esigenza di integrarsi, socializzare, autopromuoversi,* ***se si vuole che abbiano, negli atenei della penisola, un'esperienza eccellente****.*
>
> *(Iannuzzi, 2022)*

The abstract that this paper focuses on addresses teaching/learning Italian as an L2. In order to explore possible applications of MT and corpus analysis in EAP, the official translation into English of the abstract, as well as its automated translation and corpus-based post-editing are reported in the tables that follow. To foster comprehensibility, the above text is divided into several paragraphs, where the words and phrases in bold are investigated. The text of Table 1 above is translated automatically via DeepL (an MT platform) and by using Notebook LM (a language model tool). More precisely, the text inputted in DeepL is translated by selecting the British English language. On the other hand, Notebook LM is provided with 50 documents (that is the maximum number of files permitted by the LM system) on the basis of which the tool performs the tasks required. The 50 files uploaded to the platform are papers written in English dealing with teaching English as a lingua franca, as an L2, and/or for academic purposes. The prompt written in Notebook LM is the following one:

Prompt: ”On the basis of the terminology contained in the documents uploaded, translate the following text into English: [text to translate]”.

Post-editing, as mentioned, is performed by analysing the Anthology Reference Corpus available on Sketch Engine. Therefore, words, word pairs and collocations are searched for. When necessary, the online Hoepli bilingual dictionary is consulted, as well as the web as corpus by means of advanced search techniques. These are carried out via the Google search engine.

4 Analysis

This section describes in detail the corpus-driven analysis of the ARC corpus with the aim of post-editing the automated target texts. By doing so, it shows whether and how corpus consultation can be integrated into the automated translation process.

The tables that follow report the source text (first row), the official translation into English (second row), the MT- and LM-driven translations (third and fourth rows), and the corpus-based editing (fifth and last row). Alternative target words or phrases are reported in square brackets in the fifth row.

The first part of the paper is the title, i.e., “*nella classe a microlingue differenziate. progettazione didattica per l’italiano lingua seconda, di studio e di specializzazione disciplinare*”. As can be noticed, the title starts with prepositional fronting (*nella classe*, back-translation “in the classroom”) and it features no verb. See Table 3 for the official, automatic, and corpus-based translations of the title.

Table 2. Source text (the title of the paper), official translation into English, MT/LM-driven translations, and corpus-based post-editing

Source text	*Nella 1.**classe a microlingue differenziate**. 2.**progettazione didattica** per l’italiano lingua seconda, di studio e di 3.specializzazione disciplinare.*
Official target text	Italian as a second language and as a language for academic and 3.**specific purposes**. Disciplinary specialisation(s) and 2.**language curriculum design** in 1.**higher education contexts**.
MT (DeepL)	In the 1.**differentiated micro-language class**. 2.**Teaching design** for Italian as a second language, study and 3.**subject specialisation**.
Notebook LM	In 1.**class with differentiated micro-languages**. 2.**didactic planning** for italian [*sic*] as a second, study, and 3.**disciplinary specialization** language.
Corpus-based post-editing	2.**Learning [Curriculum] design** for Italian as an L2, and as a language for academic and 3.**specific purposes** in the 1.**multi-language [multilanguage] classroom**.

As far as *classe a microlingue differenziate* is concerned, the official translation reports "in higher education contexts", which is not a literal translation but a functional one, given that it explains the source words. MT features "in the differentiated micro-language class", whereas LM proposes "class with differentiated micro-languages" (see Table 2, point 1). It is evident that the LM-sourced target phrase lacks a determiner, i.e., "the". Corpus consultation may suggest alternative renditions or improve the ones shown in Table 2. By writing "microlanguage class*" in the search field of SE, it is possible to explore whether there are any concordances (i.e., hits or results) with "microlanguage" followed by any word starting with "class" (e.g., "class" or "classroom"). The asterisk, in fact, functions as a wildcard character in SE; therefore, it replaces any alphanumerical or non-alphanumerical character. Unfortunately, this query generates no hits. The same occurs if "micro-language class*", "multilanguage class*", or "multi-language class*" are searched for (one search at a time). As no consistent result is obtained from the corpus, the web can be consulted via advanced search techniques. For example, by querying "*multilanguage|microlanguage classroom|class*" *site:.ac.uk* (where the "|" symbol represents the Boolean OR operator), two hits are found (i.e., "multilanguage classroom" and "multilanguage class", respectively). By querying "*multilanguage|microlanguage classroom|class*" *site:.edu*, three results are obtained with the phrase "multilanguage/multi-language classroom". Therefore, *classe a microlingue differenziate* could be rendered as "multilanguage classroom" or "multi-language classroom" (see the last row of Table 3, point 1).

The next phrase to focus on is *progettazione didattica*, rendered as "language curriculum design" in the official translation and as "teaching design" or "didactic planning" in the automatically generated target texts.

By searching for "teach*|learn*|didact*|language*|curriculum" together with the lemma "design" or "planning" in the ARC corpus, the following words come to the fore: "language design", "curriculum design" "learning design" and "learner design". There are no instances of "didactic planning" (as suggested by LM). Hence, the translations of *progettazione didattica* can be "learning design" and/or "curriculum design" (see the last row of Table 3, point 2).

The last phrase to address in the title is *specializzazione disciplinare*. The official translation contains "specific purposes", which is adequate. By contrast, MT and LM suggest "subject/disciplinary specialisation" (see Table 3, point 3). By querying "specific purposes" in the corpus, many hits are found, where "English for specific purposes" and "language for specific purposes" prevail. By searching for "subject specialization" or "disciplinary specialisation" (or "specialisation"), only one hit is retrieved with "subject specializations". However, the phrase obtained does not address the teaching or learning of languages, as the following phrase shows: "subject specializations may include the social sciences, economics, the fine arts". Therefore, *specializzazione disciplinare* can be best rendered as "specific purposes" (see the last row of Table 3, point 3). Table 3 above shows the post-edited target text with a different word order to improve readability.

As a whole, the corpus-based translation or post-editing process highlights that word searches can be time-consuming and that users must be accustomed with search syntax. Nonetheless, the results provided are satisfactory, and translators can reach high levels of translation quality. Indeed, corpus analysis allows the retrieval of collocations and word usages in context, thereby helping select the best appropriate language options. Table 3 below reports the first paragraph of the abstract.

Table 3. Source text (first paragraph), official translation into English, MT/LM-driven translations, and corpus-based post-editing

Source text	*Questo articolo 1.**argomenta** la possibilità di 2.**contemperare** l'apprendimento dell'3.**italiano generale** con quello dell'italiano come lingua dello studio e di 4.**microlingue** specialistiche nella formazione di studentesse e 5.**studenti universitari di madrelingua non italiana** in corsi 6.**di livello A2 e B1** del Quadro comune europeo di riferimento per la conoscenza delle lingue.*
Official target text	This article 1.**discusses** the possibility of 2.**including** 3.**Italian for everyday purposes** and as a 4.**language** for academic and specific purposes in courses 6.**aimed at achieving levels A2 and B1** of the *Common European Framework of Reference for Languages*, organised by Italian universities for 5.s**tudents whose first language is not Italian**.
MT (DeepL)	This article 1.**discusses** the possibility of 2.**balancing** the learning of 3.**general Italian** with that of Italian as a language of study and of specialised 4.**micro-languages** in the training of 5.**non-Italian mother-tongue university students** in courses 6.**at level A2 and B1** of the Common European Framework of Reference for Languages.
Notebook LM	This article 1.**discusses** the possibility of 2.**balancing** the learning of 3.**general Italian** with that of Italian as a language for study and specialized 4.**micro-languages** in the training of 5.**university students whose native language is not Italian**, in courses 6.**at levels A2 and B1** of the Common European Framework of Reference for Languages.
Corpus-based post-editing	This paper 1.**explores** [investigates, describes, illustrates, addresses] the possibility of 2.**coupling** the learning of 3.**general Italian with** Italian for academic purposes and of domain-specific 4.**microlanguages** in the training of 5.**non-native Italian students** [of non-native Italian learners; of university students whose first language is not Italian; of university students whose native language is not Italian] in courses 6.**at A2 and B1 level** according to the Common European Framework of Reference for Languages.

The verb (in the 3rd person singular) *argomenta* is translated as “discusses” in both the official and automated target texts (see Table 3, point 1). By searching for “this article discusses” in the corpus, only 4 hits are obtained. If the expression “this * discusses” is queried, it is possible to retrieve words (e.g., nouns) preceding “discusses”. In this case, 317 results are gathered where the word “paper” prevails. By querying “paper|article” together with any verb in first right position, “paper” appears as more frequent than “article”. In addition, verbs such as “explores”, “investigates”, “describes”, “proposes”, “illustrates”, and “addresses” appear. Therefore, a possible translation of *questo articolo argomenta* is “this paper explores” (see the last row of Table 3, point 1). The second verb to focus on is *contemperare*, which is rendered as “include” in the official translation and “balance” by MT/LM (see Table 3, point 2). By querying “include|balance” in the corpus together with the words “learning” or “learn”, no interesting results are found (sample phrases: “include active learning", or "learning to balance the trade-off between them"). The Hoepli dictionary translates *contemperare* as "mingle together" and "blend". By searching for "blend|mingle" with the lemma "learn" in the corpus, only three (unrelated) hits are obtained (i.e., "complements the face-to-face course blended learning"; "a unique blend of machine learning", and "and blended with learn weights"). By drawing on MT and LM output (which propose the target phrase "balance the learning of"), it would be interesting to query the verbs preceding the phrase "the learning of". In this way, the following collocates are retrieved: "couple", "influence", and "facilitate". Hence, a translation of *contemperare* con in the given context can be "coupling $(\cdots)$ with" (see the last row of Table 3, point 2).

As regards *italiano in generale*, MT and LM outputs seem more accurate, as "general English" produces several hits in the corpus, whereas "English for everyday purposes" (similar to the official translation) does not appear in the corpus (see Table 3, point 3). Therefore "general Italian" is probably the best translation option. By searching for "general Italian" in Google search string, in fact, over 220,000 results are obtained. Amongst others, such nomenclature is used by universities and Italian language institutes.

The word "microlanguages", suggested by MT and LM to translate *microlingue*, produces a few hits in the corpus (one featuring the expression "domain-specific microlanguage") (see Table 3, point 4). Conversely, the word "language" (proposed in the official English version of the abstract) is too generic. As regards *microlingue specialistiche,* the expressions "specialized microlanguages" or "specialised microlanguages" (proposed by MT/LM) do not appear in the corpus. However, the noun phrase "domain-specific microlanguage" (as already noticed) can be considered as a perfect equivalent of the source phrase.

The expression *studenti universitari di madrelingua non italiana* is translated as "students whose first language is not Italian" in the official translation; "non-Italian mother-tongue university students" by DeepL, and "university students whose native language is not Italian" by Notebook LM (see Table 3, point 5). By searching for "whose first language" in the corpus, one interesting concordance is obtained, such as "children whose first language is not English". By investigating "whose native language is not", some instances are found. By querying "non-native", several results are obtained, such as "non-native English speakers", "non-native learners of English", and "non-native speakers of English". On the contrary, the expression "non-English mother" (as suggested by MT) generates no hits. Therefore, the translation of the source phrase can be "non-native Italian students", "non-native learners of Italian", or (as appearing in the LM-generated target text) "students whose native language is not Italian" (see the last row of Table 3, point 5).

The last phrase to address in the first paragraph of the abstract is *di livello A2 e B1*, which is translated as "aimed at achieving levels A2 and B1" in the original English version and as "at level(s) A2 and B1" by MT/LM tools. By querying "A2|B1|B2" with the lemma "course" or "level" in the corpus, it is possible to read "at B2 level" or "for B1 level". By searching for "Common European Framework", the following phrases come to the fore: "at around B2 level *in* the Common European Framework"; "level that is *based on* the Common European Framework"; "level A1 $(\cdots)$ from the Common European Framework", and "all language levels *according to* the Common European Framework" [emphasis added].

Therefore, corpus evidence produces several prepositions or particles which can be placed before “Common European Framework”. In light of these results, a translation option of the source phrase could be the following one: “at A2 and B1 level according to the Common European Framework of Reference for Languages” (see the last row of Table 3, point 6).

The corpus analysis carried out foregrounds the importance of searching for translation options through the lenses of lexical, semantic, and syntactic data. The use of corpora for translation and post-editing purposes relies on a multifaceted linguistic approach which considers and embraces language as a complex phenomenon, thereby going beyond algorithms and probabilistic model predictions deployed in automated translation processes. The next table reports the second part of the abstract.

Table 4. Source text (second paragraph), official translation into English, MT/LM-driven translations, and corpus-based post-editing

Source text	*1.**Dati un monte ore** e una classe (quest’ultima con 2.**determinati livelli di conoscenza in ingresso, 3.provenienza geografica spesso mista, 4.afferenza dei partecipanti a diversi corsi di laurea), 5.come includere in un corso almeno un primo approccio alle microlingue scientifico-disciplinari di interesse delle studentesse/studenti?***
Official target text	5.**Is it possible to promote an initial approach to scientific-disciplinary languages** as part of a language course, 1.**given a fixed number of learning hours** and classes composed of students of 2.**mixed first languages** and 4.**disciplinary affiliations**?
MT (DeepL)	1.**Given a number of hours** and a class (the latter with 2.**certain levels of knowledge at entry**, often 3.**mixed geographical origins**, and participants’ 4.**affiliation to different degree courses**), 5.**how can a course include at least an initial approach to the scientific-disciplinary micro-languages of interest to the students**?
Notebook LM	1.**Given a set number of hours** and a class (the latter with 2.**specific entry-level knowledge**, often 3.**mixed geographical origins**, and participants 4.**belonging to different degree programs**), 5.**how can at least a first approach to the scientific-disciplinary micro-languages of interest to the students be included in a course**?
Corpus-based post-editing	5.**This paper investigates whether it is possible to develop a course that will introduce students to scientific disciplinary microlanguages** [that will present/provide a first approach to scientific disciplinary microlanguages of interest to students] in a mixed class 1.**with a set of hours**. The class being composed of 2.**students of different native [first] languages** 3.**with mixed [diverse] origins** and 4.**different academic [university] courses [different degree programs]**.

As can be seen from Table 4 above, the official translation into English does not follow the same word order of the source text. Conversely, as could be expected, the automatically translated texts follow the same word order of the source text.

It is now interesting to investigate how the long question in 5) of Table 4 above can be rendered in English via corpus consultation. By searching for “is it possible”, 135 hits are obtained, whereas by querying “it is possible”, almost 5,000 results are retrieved. Therefore, it is apparent that statements, not questions, are generally opted for. Also, if “it is possible” is searched for together with the lemma "article" or "paper", interesting patterns come to the fore, such as "the claim of this paper is: it is possible $(\cdots)$"; "the main purpose of this paper is to show that it is possible"; "our aim in this paper is to explore whether it is possible to", and "in this paper we investigate whether it is possible to".

As regards the verb *includere*, the official text in English suggests "promote", whereas MT/LM produce "include(d)" (see Table 4, point 5). By looking for "a course" collocating with any verb to the left and to the right, the following phrases are obtained: "to develop a course of study well suited to training students for"; "a course that promises to show"; "a course that involves students learning to implement a $(\cdots)$"; "a course could present tedious work"; "a course will introduce them to programming"; "a course organized around teams of graduates", and "the paper describes a course that makes use of". By querying "a course" followed by "that" or "which", the following verb phrases are also noticed: "provides students with"; "is focussed on"; "covers a substantial amount of theory"; "presumes background knowledge they lack"; "focuses on", and "aims primarily at". Therefore, plenty of verbs and verb phrases come to the fore.

As far as *un primo approccio a* is concerned, the official target text and MT propose "an initial approach to", whereas LM suggests "a first approach to" (see Table 4, point 5). The phrase "initial approach" generates 15 hits in the corpus, and the verbs preceding it are "suggest", "present", "introduce" and "evaluate". Conversely, the phrase "first approach" produces 57 results and it collocates with verbs such as "describe", "present", and "provide".

The adjectival phrase *scientifico-disciplinari* is translated as "scientific-disciplinary" in all target texts (see Table 4, still point 5). However, by exploring "disciplinary" with the lemma "scientific", no hits are retrieved from the corpus. By carrying out targeted web searches, it is possible to read the noun phrase "scientific disciplinary" with no hyphen (search queries: "*scientific-disciplinary*" *site:.edu* and "*scientific-disciplinary*" *site:.ac.uk*).

Finally, point 5 in Table 4 mentions *di interesse delle studentesse/studenti*, whose rendition is omitted in the official translation, although it is translated as "of interest to the students" in the automated target texts. If the prepositional phrase "of interest to" is searched for in the corpus together with the lemma "student", the following phrases are retrieved: "of interest to the student" and "of interest to ESL students". Therefore, "of interest to the students" can be satisfactory. Nonetheless, the target phrase proposed by LM cannot go unnoticed: "how can at least a first approach to the scientific-disciplinary micro-languages of interest to the students be included in a course?". As observable, the question features syntactic discontinuities and the use of the passive form. Particularly, the subject is too long and syntactically complex (i.e., "a first approach to the scientific-disciplinary micro-languages of interest to the students).

In light of the above, the question that is posed in the source text (Table 4, point 5), is rendered as follows: "this paper investigates whether it is possible to develop a course that will introduce students to scientific disciplinary microlanguages" or "this paper explores whether it is possible to develop a course that will present an initial approach to scientific disciplinary microlanguages of interest to students" (see last row of Table 4, point 5).

It is now possible to analyse the past participle phrase which sets the start of the paragraph. As far as *dati un monte ore* is concerned, the official translation reads "given a fixed number of learning hours", whereas MT- and LM-driven outputs suggest "given a (set) number of hours". If "number of hours" is queried in the corpus, a few hits are retrieved, whereas by searching for "number of learning hours", no results are found. Listing the adjectives preceding the word "hours" does not yield any relevant results (e.g., "academic hours"). However, by investigating "hours" collocating with "set", "fixed" or "limited", the phrase "a set of 20 hours" surfaces. Hence, *monte ore* could be rendered as "a set of hours" (see Table 4, point 1).

The expression *determinati livelli di conoscenza in ingresso* is translated as "mixed first languages", whereas MT proposes "certain levels of knowledge at entry" and LM "specific entry-level knowledge" (see Table 4, point 2). The corpus generates several results with "entry-level" (e.g., "entry-level categories" or "entry-level course"). There are, however, no results with "entry-level knowledge". By listing the collocations preceding "first language*", the words "different", "many" and "various" come to the fore. If "native language*" is queried, the modifiers "different", "several" and "other" are found. Therefore, the source phrase can be rendered as "different native/first languages" (see the last row of Table 4, point 2).

The phrase *provenienza geografica spesso mista* is unaddressed in the official translation, whereas it is rendered as "mixed geographical origins" by the MT and LM tools (see Table 4, point 3). By searching for "mixed geogr*", only one hit is found, which is unrelated. Querying "geographical origin" generates three hits, although there are no modifiers translating *mista* ("mixed"). By searching for the adjectives preceding "origin", the following phrases appear: "nonnative speakers of different origins"; "diverse origin"; "disparate origin"; "with various (linguistic) origins"; "with mixed origins", and "with multiple origins". Therefore, the source phrase can be rendered as "with mixed / diverse / different origins" (see the last row of Table 4, point 3).

Finally, *afferenza dei partecipanti a diversi corsi di laurea* appears as "(mixed) disciplinary affiliations" in the official translation, as "affiliation to different degree courses" in DeepL's target text and as "belonging to different degree programs" in Notebook LM's output (see Table 4, point 4). The noun phrase "disciplinary affiliation" generates no hits in the corpus, as well as "degree course(s)". The noun phrase "degree program(s)", conversely, produces a few hits. By searching for the adjectives and nouns collocating with "course" in the corpus, the following expressions are found: "academic courses" and "university course". Consequently, the source phrase can be translated as "different academic / university courses" or "different degree programs" (see the last row of Table 4, point 4).

As with the other cases, corpus-based post-editing reveals to be a complex but satisfactory process thanks to which several translation options can be sourced. Table 5 below shows the third and last paragraph. With regard to *attività didattiche*, the noun phrase "learning activities" (written in the English translation) generates 37 hits in the corpus, whereas "teaching activities" (visible in the MT and LM texts) produces no results. For these reasons, the first option is considered (see Table 5, point 1).

As far as the verb phrase *delineare questa possibilità* is concerned, all target texts propose "outlining" to render *delineare* (see Table 5, point 2).

Table 5. Source text (third paragraph), official translation into English, MT/LM-driven translations, and corpus-based post-editing

Source text	Il lavoro di progettazione può rispondere a questa domanda concentrando la propria azione sul design del curricolo e di alcune 1.**attività didattiche** specifiche. Nel 2.**delineare questa possibilità**, questo articolo intende 3.**fornire uno spunto utile** a quelle/i insegnanti e coordinatrici/ori che 4.**si trovano a operare entro precisi vincoli organizzativi**, senza voler rinunciare a un'offerta didattica che tenga conto delle plurime dimensioni in cui studentesse e studenti di madrelingua non italiana hanno esigenza di integrarsi, socializzare, autopromuoversi, 5.**se si vuole che abbiano, negli atenei della penisola, un'esperienza eccellente**.
Official target text	The affirmative answer to this question lies in the design of the curriculum and specific 1.**learning activities**. By 2.**outlining these**, this article aims to 3.**facilitate the work** of those teachers and coordinators who 4.**find themselves operating within given organisational limits**, and trying to take into account the multiple dimensions in which students whose first language is not Italian need to integrate, socialise, and succeed in their studies, 5.**if we want them to have an excellent experience in the Peninsula's universities**.
MT (DeepL)	Design work can answer this question by focusing its action on the design of the curriculum and some specific 1.**teaching activities**. In 2.**outlining this possibility**, this article intends to 3.**provide a useful cue** for those teachers and coordinators who 4.**find themselves operating within precise organisational frameworks**, without wishing to renounce a didactic offer that takes into account the multiple dimensions in which non-Italian mother-tongue students need to integrate, socialise, and self-promote, 5.**if they are to have an excellent experience in the universities of the peninsula**.
Notebook LM	The planning work can answer this question by focusing its action on the design of the curriculum and some specific 1.**teaching activities**. In 2.**outlining this possibility**, this article aims to 3.**provide a useful starting point** for those teachers and coordinators who 4.**find themselves operating within precise organizational frameworks**, without wanting to give up a didactic offer that takes into account the multiple dimensions in which students of non-Italian mother tongue need to integrate, socialize, and promote themselves, 5.**if we want them to have an excellent experience in the universities of the peninsula**.
Corpus-based post-editing	The design of a specific curriculum and of 1.**learning activities** can answer this question. In 2.**investigating [suggesting, discussing, exploring; while pursuing] this possibility**, this paper aims to 3.**provide [offer / give] strong hints** to lecturers and coordinators who 4.**work in a specific organisational context** [in a certain organisational setting / within a specific organisational framework] (· · ·) 5.**and allow them to have** [help them have; if they are to have] **a rewarding [valuable / positive / successful] experience in the universities of the Peninsula [in the Peninsula's universities]**.

By searching for the verbs preceding the lemma "possibility" in the corpus, the following ones emerges: “consider”, “signal”, “discuss”, “investigate”, “illustrate”, “highlight”, “suggest”, “explore” and many others. As the source phrase starts with a preposition (i.e., *nel delineare questa possibilità*), it could be useful to search for particles preceding “possibility” in the corpus. Therefore, the lemma “possibility” is queried together with “in” and/or “while” up to the third left position. The following phrases are obtained: “in exploring the possibility of” and “while pursuing this possibility”. Acceptable translation solutions of the source phrase are “in investigating / suggesting / discussing / exploring this possibility” and “while pursuing this possibility” (see the last row of Table 5, point 2).

The expression *fornire uno spunto* utile is paraphrased as “facilitate the work” in the official English version, and translated as “provide a useful cue” or “provide a useful starting point” in the automated target texts (see Table 5, point 3). The Hoepli dictionary proposes “cue”, “hint” and “idea” as translations of *spunto*. If the verbs collocating with “cue|hint|idea” are searched for in the corpus, the following phrases are found: “give an idea”; “provides hints to”; “describe the basic idea of”; “yield better cues for”; “offer / give strong hints”, and “provide strong cues”. Therefore, a translation option of the source phrase could be “provide / offer / give strong hints” (see the last row of Table 5, point 3).

The source phrase *si trovano a operare entro precisi organizzativi* is missing the object *limiti* (its back-translation reads “find themselves operating within precise organisational”). As a matter of fact, the official English version adds the word “limits” (i.e., “find themselves operating within given organisational limits”). The automatically generated target texts integrate the source text with “frameworks” (i.e., “find themselves operating within precise organisational frameworks”) (see Table 5, point 4). As can be seen, both renditions present “find themselves”, “operating” and “within”. If “*in * organisational” and “*in * organizational” are queried in the corpus, the following phrases are retrieved: “within an organisational framework”; “in an organisational context”, and “in an organizational setting”. By searching for the adjectives preceding “organisational|organizational”, the following expressions are obtained: “certain organizational principles”; “established organizational structure”, and “specific organizational features”. Also, as the verb *operare* refers to the activities of teachers and coordinators, its rendition would be best expressed by the verb “work”. Therefore, possible translations of the source phrase are “work in a specific organisational context”; “work in a certain organisational setting”, and/or “work within a specific organisational framework” (see the last row of Table 5, point 4).

The last phrase to tackle is *se si vuole che abbiano, negli atenei della penisola, un’esperienza eccellente*, translated as “if we want them to have an excellent experience in the Peninsula’s universities” in the official translation into English. MT- and LM-driven output is as follows: “if they are / if we want them to have an excellent experience in the universities of the peninsula” (see Table 5, point 5). The expression “we want * to have” produces only four hits in the corpus, whereas “are to have” generates six results. Alternatives to these expressions could be verb phrases such as “allow (someone) to”, “help (someone)”, or “let (someone)”. The search for “allow * to” produces over 10,000 hits (e.g., “allow users to”). The lemma “help” followed by a pronoun generates over 1,000 results, whereas the lemma “let” followed by any pronoun yields more than 3,000 hits (albeit mostly in phrases containing “let us”). Therefore, the expression “allow them to” could be considered as an acceptable translation of *se si vuole che* (see the last row of Table 5, point 5).

The phrase *abbiano (· · ·) un'esperienza eccellente* is rendered as "have an excellent experience" in all target texts (see Table 5, point 5). By searching for the verbs preceding "experience", "have" and "gain" come to the fore in the following expression: "was a valuable experience". By listing the adjectives collocating with "experience", these words are obtained and can be considered as translations of *eccellente:* "fantastic", "great", "good", "positive", "productive", "rewarding", "thorough", and "successful". Therefore, the source phrase is rendered as "have a rewarding / valuable / positive / successful experience".

The prepositional phrase *negli atenei della penisola* is translated as "in the Peninsula's universities" in the official target text and as "in the universities of the peninsula" in the automatic texts (see Table 5, still point 5). By querying either "in the * universit*" or in the universit* of", 22 hits are found. Given the equal results, it may be useful to carry out targeted web searches. The string "*the peninsula's university|universities*" *site:.edu* generates 5 hits, whereas "*the peninsula's university|universities*" *site:.ac.uk* produces only 1 result (with no apostrophe after "peninsula"). By contrast, the string "*the university|universities of the peninsula*" *site:.edu* retrieves 4 hits, and "*the university|universities of the peninsula*" *site:.ac.uk* generates 2. In view of such results, both target phrases can be considered as acceptable. Therefore, the source phrase can either be "of the Peninsula's universities" or "of the universities of the Peninsula" (see the last row of Table 5, point 5).

5 Discussion

On the basis of the analyses carried out, it can be inferred that MT- and LM-driven output is helpful to suggest translation options. However, post-editing must be carried out mindfully and terms must often be disambiguated. Scholars, in fact, have recently shifted their attention to machine-driven literacy and how to become aware of the complementarity of human and artificial intelligence (Ehrensberger-Dow et al. 2023).

In this regard, corpus analysis proved to be satisfactory as it helped fine-tune both the translation and the post-editing processes. For example, the expression "general Italian" (Table 3), as proposed by machines, is more frequent than "Italian for everyday purposes" (written in the official English version of the abstract). This was ascertained thanks to corpus analysis.

Other times, both the official translation and MT/LM did not suggest frequent language combinations, as in the target phrase "this article discusses" (Table 3). In this case, the corpus-sourced phrase "this paper explores" is proved to be more frequently used in academic settings.

Also, as noticed throughout the analyses, the official translation into English explained source terms or phrases rather than translating them. An example is *microlingue specialistiche*, rendered as "language for academic and specific purposes" (Table 3). In this regard, corpus-based analysis helped retrieve more adherent renderings, such as "domain-specific microlanguages", whereas the MT/LM-proposed phrase ("specialised microlanguages" or "specialized microlanguages") did not frequently appear in the academic corpus.

Another interesting example is the expression *di livello A2 e B1* (Table 3), which was rendered as "aimed at achieving levels A2 and B1" in the official English translation. Corpus consultation has allowed for simplification by translating the source phrase as "at A2 and B1 level". The same can be said of *studenti universitari di madrelingua non italiana*. The official translation addressed this expression by proposing a phrase that was not featured in the corpus (i.e., "students whose first language is not Italian"), whereas MT rendering was rather intricate (i.e., "non-Italian mother-tongue university students"). In this case, corpus analysis helped composing a more natural-sounding phrase, i.e., "non-native Italian students", which is probably more frequent.

The analysis carried out in Table 4 corroborates that corpus consultation helps produce target language patterns that comply with the academic writing style. The question posed at the end of the paragraph, for example, was turned into a statement and placed at the beginning. In this way, the reader is given details on what the paper explores from the beginning of the paragraph. In addition, corpus analysis foregrounded varied translation options, such as the verbs "present" and "provide" to translate *includere* (rendered as "promote" in the official translation and "include" in MT/LM). Also, there were expressions in the official translation that could be considered as uncommon or infrequent in academic writing. Some of these were, for example, "learning hours", or "disciplinary affiliations". Thanks to corpus analysis, such phrases were changed into "a set of hours" and "academic / university courses", respectively.

Finally, Table 5 corroborates that the functional translation strategies followed in the official translation paraphrased source terms or expressions rather than translating them. An example is the verb phrase "facilitate the work" which rephrases and simplifies *fornire uno spunto utile*. Moreover, as already noticed, corpus analysis brought to the fore a variety of valid alternative translations, as in the case of *delineare questa possibilità*, where equivalents of the verb *delineare* were manifold (i.e., "investigating", "suggesting", "discussing", "exploring" and "pursuing"). The same occurred to *esperienza eccellente*, where *eccellente* could be translated with a wide range of modifiers.

For these reasons, corpus consultation is a reliable post-editing tool which helps enhance the quality of (machine) translation outputs.

6 Conclusion

On the basis of the analyses carried out and of the results obtained, the automated translations of the source text combined with corpus-based post-editing allows users to develop critical thinking by stimulating cognitive processes. Corpus consultation, in fact, helps retrieve and assess a variety of translation alternatives and notice language patterning. Additionally, it enables users to explore word usages in contexts and collocations. Corpus evidence provides samples of recurrent vs less recurrent patterns of language. In this way, translation skills are enhanced.

In the case in point, the official translation into English was useful as it could be a teaching tool which stimulates reflections on the translation choices adopted. Furthermore, it helps verify whether and to what extent MT- and LM-driven outputs can be modified.

As a whole, the paper highlighted the fact that automated translations often need reformulating and cannot be considered as final target texts. For these reasons, however, they can be exploited in the (corpus-based) translator training classroom.

The implications for language instructors and the applications in the field of Translation

Studies are manifold. Firstly, trainers can help learners focus on the retrieval of collocates from targeted corpora and let them discover language patterns and nuances. Secondly, they may prompt learners to analyse word usages in context, thus deriving important syntactical and semantic data. Thirdly, with the help of bilingual dictionaries, students can be taught how to disambiguate words in context thanks to corpus consultation. In this way, they will get a grasp of the varied usages and meanings that terms can be imbued with not only at phrase level, but also on the basis of the field or genre focused on. Corpora have become important tools which facilitate language analysis and stimulate reflections on linguistic structures and meanings. They support grammatical and semantic choices, thus developing critical thinking.

The first question that this paper wished to address was "How can the automatic translation of an academic abstract be used in translator training to foster both writing and translation skills?". This paper showed how an automatically translated academic abstract can be used to improve translation at academic level. For this reason, the methodology followed in this paper can be replicated in translator training. Students can be exposed to machine-generated outputs and prompted to post-edit it via corpus consultation and/or by targeted web searches. The second question asked whether and how corpus-based post-editing activities of machine-translated texts could be integrated in the EAP classroom. As demonstrated, corpus-driven data allows users to successfully post-edit target text, notice collocations, become acquainted with word usages in contexts, and find language patterning. The corpus-driven searches displayed in this paper can be replicated to engage students in corpus-based activities. Therefore, the answer to the third and last question ("Does the corpus produce satisfactory evidence to be successfully integrated in the automated translation process?") is affirmative.

The limits of this paper lie in the fact that only one abstract was considered. A wider variety of source and target texts may have led to more insightful results. Therefore, these initial findings could be corroborated or challenged by other studies. Another limitation is based on the fact that evidence of (academic) language patterning was attested on one corpus only (as well as on targeted web-driven searches). Consulting more academic corpora might have generated different results.

Future research could apply the methodology and text analyses proposed in this paper in classroom observation studies and verify how translation skills are enhanced in the short and/or in the long run. Moreover, further research could envisage the consultation of various academic corpora or of an *ad hoc* DIY corpus.

Riferimenti bibliografici

Biber, D. and Conrad, S. (2009). *Real Grammar – A Corpus-based Approach to English*. Pearson Longman, New York.

Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., and Fine, J. (1988). Reading english for specialized purposes: Discourse analysis and the use of standard informants. In Carrell, P., Devine, J., and Eskey, D., editors, *Interactive Approaches to Second Language Reading*, pages 152–167. Cambridge University Press, Cambridge.

Crème, P. and Lea, M. R. (2008). *Writing at University: A Guide for Students*. McGraw-Hill House, Maidenhead, 3rd edition.

DeepL (2025). Deepl translator. `https://www.deepl.com/it/translator`.

Dunleavy, P. (2003). *Authoring a PhD: How to Plan, Draft, Write and Finish a Doctoral Thesis or Dissertation*. Palgrave Macmillan, New York.

Ehrensberger-Dow, M., Delorme Benites, A., and Lehr, C. (2023). A new role for translators and trainers: Mt literacy consultants. *The Interpreter and Translator Trainer*, 17(3):393–411.

Flowerdew, L. (2015). Corpus-based research and pedagogy in eap: From lexis to genre. *Language Teaching*, 48(1):99–116.

Fortanet-Gómez, I. (2006). Interaction in academic spoken english: The use of 'i' and 'you' in the micase. In Macià, E. A., Cervera, A. S., and Ramos, C. R., editors, *Information Technology in Languages for Specific Purposes. Educational Linguistics, 7*, pages 35–51. Springer, Boston, MA.

Giampieri, P. (2017). *Academic English*. De Agostini, Torino.

Google (2025). Notebook lm. `https://notebooklm.google/`.

Götz, A. (2015). Magyarés angol abstraktok retorikai szerkezetének elemzése [analysis of the rhetorical structure of hungarian and english abstracts]. *Fordítástudomány [Translation Studies]*, 17(2):88–116.

Hoepli Dizionari (2025). Hoepli dictionaries. `https://dizionari.repubblica.it`.

Hyland, K. (1995). The author in the text: Hedging scientific writing. *Hong Kong Papers in Linguistics and Language Teaching*, 18:33–42.

Hyland, K. (2007). Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics*, 28(2):266–285.

Hyland, K. (2011). Disciplines and discourses: Social interactions in the construction of knowledge. In Starke-Meyerring, D., Paré, A., Artemeva, N., Horne, M., and Yousoubova, L., editors, *Writing in Knowledge Societies*, pages 193–214. Parlor Press and The WAC Clearinghouse, West Lafayette, IN.

Hyland, K. (2015). Corpora and written academic english. In Biber, D. and Reppen, R., editors, *The Cambridge Handbook of English Corpus Linguistics*, pages 292–308. Cambridge University Press, Cambridge.

Hyland, K. and Hamp-Lyons, L. (2002). Eap: Issues and directions. *Journal of English for Academic Purposes*, 1:1–12.

Iannuzzi, G. (2022). Nella classe a microlingue differenziate. progettazione didattica per l'italiano lingua seconda, di studio e di specializzazione disciplinare. *Italiano Lin-*

guaDue, (2):703–713.

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology*.

Lee, D. Y. W. and Swales, J. M. (2006). A corpus-based eap course for nns doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes*, 25(1):56–75.

Lee, S. M. (2023). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning*, 36(1-2):103–125.

Lexical Computing (2025). Sketch engine. `http://www.sketchengine.eu`.

López-Arroyo, B. and Méndez-Cendón, B. (2007). Describing phraseological devices in medical abstracts: An english/spanish contrastive analysis. *Meta*, 52(3):503–516.

Mauranen, A. (2003). The corpus of english as lingua franca in academic settings. *TESOL Quarterly*, 37(3):513–527.

Mauranen, A. (2012). *Exploring ELF: Academic English Shaped by Non-native Speakers*. Cambridge University Press, Cambridge.

Molino, A. (2014). Vague lexis in spoken academic english and in advanced corpus-based learner's dictionaries. In Molino, A. and Zanotti, S., editors, *Observing Norms, Observing Usage: Lexis in Dictionaries and in the Media*, pages 219–238. Peter Lang, Berlin.

Morton, R. (1999). Abstracts as authentic material for eap classes. *ELT Journal*, 53(3):177–182.

O'Brien, S., Simard, M., and Goulet, M. J. (2018). Machine translation and self-post-editing for academic writing support: Quality explorations. In Moorkens, J., Castilho, S., Gasperi, F., and Doherty, S., editors, *Translation Quality Assessment: From Principles to Practice*, Machine Translation Series, pages 237–262. Springer International Publishing, Cham.

O'Keeffe, A., McCarthy, M., and Carter, R. (2007). *From Classroom to Corpus – Language Use and Language Teaching*. Cambridge University Press, Cambridge.

Olkhovska, A. and Frolova, I. (2020). Using machine translation engines in the classroom: A survey of translation students' performance. *Advanced Education*, 15:47–55.

Randaccio, M. (2013). Writing skills: theory and practice. *QuaderniCIRD*, 7:51–74.

Scarpa, F. (2020). *Research and Professional Practice in Specialised Translation*. Palgrave Macmillan, London.

Shei, C. C. C. (2005). Translation commentary: A happy medium between translation curriculum and eap. *System*, 33(2):309–325.

Siegel, J. (2023). Translanguaging options for note-taking in eap and emi. *ELT Journal*, 77(1):42–51.

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford University Press, Oxford.

Swales, J. M. (2006). Corpus linguistics and english for academic purposes. In Arnó Macià, E., Soler Cervera, A., and Ramos Rueda, C., editors, *Information Technology in Languages for Specific Purposes: Issues and Prospects*, pages 19–33. Springer, New York.