Part 1

Applicative and Scientific Context
Leveraging Comparable Corpora for Computer-assisted Translation

1.1. Introduction

This chapter starts with a historical approach to computer-assisted translation (section 1.2): we will retrace the beginnings of machine translation and explain how computer-assisted translation has developed so far, with the recent appearance of the issue of comparable-corpus leveraging. Section 1.3 explains the current techniques to extract bilingual lexicons from comparable corpora. We provide an overview of the typical performances, and discuss the limitations of these techniques. Section 1.4 describes the prototyping of the computer-assisted translation (CAT) tool meant for comparable corpora and based on the techniques described in section 1.3.

1.2. From the beginnings of machine translation to comparable corpora processing

1.2.1. The dawn of machine translation

From the beginning, scientific research in computer science has tried to use the machine to accelerate and replace human translation. According to [HUT 05], it was in the United States, between 1959 and 1966, that the first research in machine translation was carried out. Here, machine translation (MT) refers to the translation of a text by a machine without any human intervention. Until 1966, several research groups were created, and two types of approaches could be identified:
– On the one hand, there were the pragmatic approaches combining statistical information with trial-and-error development methods\(^1\) and whose goal was to create an operational system as quickly as possible (University of Washington, Rand Corporation and University of Georgetown). This research applied the direct translation method\(^2\) and this gave rise to the first generation of machine translation systems.

– On the other hand, theoretic approaches emerged involving fundamental linguistics and considering research in the long term (MIT, Cambridge Research Language Unit). These projects were more theoretical and created the first versions of interlingual systems.\(^3\)

In 1966, a report from the Automatic Language Processing Advisory Committee [ALP 66], which assesses machine translation purely based on the needs of the American government – i.e. the translation of Russian scientific documents – announced that after several years of research, it was not possible to obtain a translation that was entirely carried out by a computer and of human quality. Only postedition would allow us to reach a good quality of translation.\(^4\) Yet the point of postedition is not self-evident. A study mentioned in the appendix of this book points out that “most translators found postediting tedious and even frustrating”, but many found “the output served as an aid... particularly with regard to technical terms” [HUT 96].

Although the study does not allow us to come to a conclusion on the point of postedition in relation to fully manual translation (out of 22 translators, eight find postedition easier, eight others find it harder and six were undecided), the report mostly highlights the negative aspects, quoting one of the translators:

I found that I spend at least as much time in editing as if I had carried out the entire translation from the start. Even at that, I doubted if the edited translation reads as smoothly as one which I would have started from scratch. [HUT 96]

---

1 Several heuristic rules were implemented and tested on data until the result obtained was considered satisfactory.

2 A translation strategy, which does not involve any mediating processing layer: the very first translators used this approach by tokenizing a text into words, neutralizing inflections, looking for the translation of words in a bilingual dictionary, and after all this the translated words were reordered following several rules. Therefore, there was no syntactic or semantic analysis.

3 The interlingual method analyzes the source text so as to generate an abstract semantic representation of it, which is totally language-independent. The target text is then generated based on this representation. The generation module in the target language can only access the interlingual representation.

4 MT “presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing” – quoted in [HUT 96].
The report quotes remarks made by V. Yngve – the head of the machine translation research project at MIT – who claimed that MT “serves no useful purpose without postediting, and that with postediting the over-all process is slow and probably uneconomical” [HUT 96].

The report concludes on the fact that machine translation research is essential from the point of view of scientific progress, it however has a limited interest from an economic point of view. Thus funding was cut in the United States. However, research carried on in Europe (EUROTRA research project) and in Canada. This research was the source of the TAUM system, for example, (translation of weather reports from French to English) and of the translation software SYSTRAN.

1.2.2. The development of computer-assisted translation

While it signaled the end of public funding for machine translation research in the United States, the ALPAC report encouraged the pursuit of a more realistic goal for computer-assisted translation. The report praised the glossaries generated by the German army’s translation agency as well as the terminology base of the European Coal and Steal Community – a resource which foregrounded EURODACOM and IATE – and came to the conclusion that these resources were a real help to translation. The final recommendations clearly encouraged the development of CAT, especially in the leveraging of glossaries initially created for machine translation.

At that point, a whole range of tools intended to help the translator in his/her work rather than replace him/her started to be developed. The first terminology management programs appeared in the 1960s [HUT 05] and evolved into multilingual terminology databases such as TERMiUM or UNTERTM. Bilingual concordancers are also of invaluable help: they allow the translator to access the word or term’s context and compare the translation of the contexts in the target language. According to [SOM 05], the rise in computer-assisted translation happened in the seventies with the creation of translation memory software, which allows the translator to recycle past translations: when a translator has to translate a new sentence, the software scans the memory for similar previously translated sentences, and when it finds any, suggests the previous translation as translation model. The

---

5 “Machine-aided translation may be an important avenue toward better, quicker and cheaper translation” quoted in [HUT 96].
6 “research should be supported on: [...] 2. means for speeding up the human translation process; [...] 6. evaluation of the relative speed and costs of various sorts of machine-aided translation; 7. adaptation of existing mechanized editing and production processes in translation; [...] 9. production of adequate reference works for the translator, including the adaptation of glossaries that now exist primarily for automatic dictionary look-up in machine translation” quoted by [HUT 96].
time saved is all the greater when the texts translated are repetitive, which is often the case in certain specialized documents such as technical manuals.

These sets of translated documents make up what we call parallel corpora\textsuperscript{7} [VER 00] and their leveraging intensified in the 1980s, allowing for a resurgence in machine translation. While the translation systems based on rules had dominated the field until then, the access to large databases of translation examples helped further the development of data-driven systems. The two paradigms arising from this turnaround are the example-base translation [NAG 84] and statistical machine translation [BRO 90], which remains the current dominant trend. The quality of machine translation is improving. Today, it generates usable results in specialized fields in which vocabulary and structures are rather repetitive. The last stronghold is general texts: machine translation offers, at best, an aid for understanding.

During the 1990s, CAT benefited from the intersecting input of machine translation and computational terminology [BOU 94, DAI 94a, ENG 95, JAC 96]. It was at that point that term alignment algorithms appeared, based on parallel corpora [DAI 94b, MEL 99, GAU 00]. The bilingual terminology lists generated are particularly useful in the case of specialized translation.

Automatic extraction and management of terminology, bilingual concordance services, pre-translation and translation memories, understanding aids: today, the translator’s workstation is a complex and highly digital environment. The language technology industry has proliferated and developed itself, generating many pieces of CAT software: TRADOS\textsuperscript{8}, WORDFAST\textsuperscript{9}, DÉJÀ VU\textsuperscript{10}, and SIMILIS\textsuperscript{11} to name just a few. The greater public is also provided for: on the one hand, Google has widened the access to immediate translation for anyone due to its GOOGLE TRANSLATE tool\textsuperscript{12} and on the other hand, open access bilingual concordance services have appeared recently on the Internet (BAB.LA\textsuperscript{13}, LINGUEE\textsuperscript{14}), and quickly become popular – for example LINGUEE reached 600,000 requests a day for is English–German version in 2008, a year after it had been created [PER 10].

\textsuperscript{7} “texts accompanied by their translation in one or more languages” [VER 00].
\textsuperscript{8} www.trados.com.
\textsuperscript{9} www.wordfast.com.
\textsuperscript{10} www.atril.com.
\textsuperscript{11} www.lingua-et-machina.com.
\textsuperscript{12} www.translate.google.com.
\textsuperscript{13} www.en.bab.la.
\textsuperscript{14} www.linguee.com.
1.2.3. Drawbacks of parallel corpora and advantages of comparable corpora

While they are useful, these technologies have a major drawback: they require the existence of a translation history. What about languages, which have few resources or emerging speciality fields? A possible solution is then to use what we refer to as comparable corpora.

There exist several definitions of comparable corpora. At one end of the spectrum is the very narrow definition given by [MCE 07]; within the framework of translation studies research. According to these authors, a comparable corpus contains texts in two or more languages, which have been gathered according to the same genre, field and sampling period criteria. Moreover, the corpora must be balanced: “comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness (McEnery, 2003:450), e.g. the same proportions of the texts of the same genres in the same domains in a range of different languages in the same sampling period. However, the subcorpora of a comparable corpus are not translations of each other. Instead, their comparability lies in their same sampling frame and similar balance [MCE 07]. At the other end of the spectrum, we encounter the definition given by [DÉJ 02], within the framework of natural language processing research, which only underlines the fact that there should be “a substantial subpart” of vocabulary in common between the texts15.

As for us, we have chosen a middle point, considering that sets of texts are comparable, if they are in two or more languages dealing with a same topic and if possible, if they have been generated within the same communication situation, so that there is a possibility of finding useful translations in them. We will only look at specialized comparable corpora, i.e. the texts generated by an expert in the field and addressed to other experts or the general public [BOW 02].

As well as being more easily available, comparable corpora also have an advantage in quality, which is emphasized by translation studies researchers. Parallel corpora are well-known for not being faithful to linguistic uses in the target language. For [MCE 07], “translated language is at best an unrepresentative special variant of the target language” [MCE 07]. For [ZAN 98], translated texts cannot represent all the linguistic possibilities of the target language and tend to reflect the idiosyncrasies of the source languages as well as those of the translator. As for [BAK 96], she

15 “Deux corpus de deux langues I et I2 sont dits comparables s’il existe une sous-partie non négligeable du vocabulaire du corpus de langue I, respectivement I2, dont la traduction se trouve dans le corpus de langue I2, respectivement I1”. Translation: two corpora in two languages I and I2 are comparable if there is a substantial subpart of the vocabulary of the corpus of language I, I2 respectively, whose translation is in the body language of I2 and I, respectively [DÉJ 02].
explains how the texts generated by a translation, like any other text, are influenced by their production context and the communication goals that they serve. Thus, they have specific characteristics, which differentiate them from “spontaneous” texts.

The term translationese is used to refer to this variation of language, which is generated in a translation situation. The existence of translationese has been widely studied and proven. Its characteristics are visible by comparing a translation corpus with a corpus of spontaneous texts covering the same topic.

[BAK 96] synthesize the results of several studies mainly based on the comparison between original texts and translations in English (newspaper articles and novels).

She highlights four characteristics:

Clarifying: clarifying is the tendency to avoid the implicit, and even to add additional information to replace the message in context. Translated texts are always longer than the source text, no matter what the translation direction or the languages are: from a lexical point of view, we notice more explanatory vocabulary (cause, reason) and connectives such as because, consequently.

Simplification: the language used is simplified. Sentences that are too long are cut up into shorter sentences. Punctuation is changed: weak punctuation marks are replaced by stronger punctuations (from comma to semi-colon to period). The translations have less lexical variety and a stronger proportion of tool words.

Standardization / conservatism: this aspect concerns the conformity or even the exaggeration of the typical characteristics of the target language, especially with regards to grammatical structures, punctuation and collocations.

Levelling out: translated texts show much less variety than spontaneous texts in numerous ways. For example, if we look at the variations of the type: token ratio (which measures the lexical variety) or of the sentence length over several texts, the variation of these characteristics is much lower for translated texts.

In the case of comparable corpora, several studies have underlined their usefulness for translation.

Two studies [FRI 97, GAV 97], mentioned by [MCE 07], estimate that specialized comparable corpora are useful in technical translation when it comes to checking translation hypotheses. [FRI 97] noticed improvements in quality, whether it is translated toward the translator’s first or second language. The fact that there is an improvement even in the case of a translation toward the first language is proof of how hard it is to approach specialized texts. Indeed, being able to use everyday language does not mean that we know the terminology or linguistic uses specific to a field, or even the notions and concepts, which they deal with.

The works of [ZAN 98] on translator training highlight three possible uses of comparable corpora:
Researching translation matches: [ZAN 98] describes an experiment on the identification of translational matches in sport newspapers, which are said to employ a large amount of figurative language. The example given is the translation of the expression salire il gradino più alto del podio (to climb on the highest step of the podium) into English: can it be translated literally or should a matching term be chosen? The corpus study of the contexts of occurrence of the Italian expression show that this expression means to win the gold medal. A study of the joint occurrence of the word podium in English texts shows that although the meaning is the same as the Italian podio, podium does not appear with the highest step to denote winning the gold medal. A literal translation would thus be a poor translation, and the chosen translation will be to win the gold medal.

Learning terminology: [ZAN 98] underlines the strong proportion of translation matches between terms that are graphically similar in medical corpora (terms with common Greek and Latin origins, for example, i.e. hépatique ↔ hepatic). He explains that the observation of the collocations of similar terms such as these can help acquire knowledge of field-specific terminology. The example given is that of the translation of biopsia epatica, which intuitively in English would be hepatic biopsy. However, the context of biopsy never mentions the expression hepatic biopsy whereas liver biopsy appears 39 times. A more in-depth study of the contexts of liver versus fegato (layman terms) and hepatic versus epatico/a (scholarly terms) show that the English and the Italian do not use layman and scholarly terms in the same way: in English, hepatic only occurs in the company of generic terms such as lesion or disease whereas in Italian, the scholarly term is used without any kind of restriction.

Exporting texts post- and pre-translation: in this case, we use comparable corpora to examine the uses specific to a field or a genre. The experiment described concerns a comparative study in the appearance of the word Mitterand in English and Italian newspapers. This study reveals that there are stylistic traditions in each language: in Italian, we tend to refer to politicians by their full name (François Mitterand) whereas in English, we use their title more often (Mr. Mitterand, President Mitterand). These uses are also different when it comes to introducing reported speech: in English, a small number of verbs is used (say and add are used in 60% of the cases) whereas in Italian, the verbs used to report speech are much more varied.

1.2.4. Difficulties of technical translation

To explain the difficulties of technical translation, we will rely on Christine Durieux’s work ([DUR 10]), which subscribes to Danica Seleskovich’s interpretative theory of translation (or theory of meaning).

At first, one may believe that specialized human translation only focuses on the acquisition of translation matches between terms (learning terminology). Yet, as [DUR 10] explains, technical translation cannot be limited to the process of
generating terminology matches. This approach is what she calls “transcoding”, which is simply the transposition into the target language of terms that are not necessarily understood. The writer believes that a good technical translation can only exist if the translator is completely at home with the notions referred to in these terms: “one does not translate a sequence of words, but a message whose meaning was first understood” 16 [DUR 10]. Thus, the translator’s work involves a dimension of self-improvement in the technical field, carried out through prior documentation enabling him or her to learn the field’s terminology in context. Durieux suggests carrying out this documentation research in educational and outreach material such as encyclopedias in which the notions are described with an easily-accessible vocabulary rather than in specialized resources such as scientific journals. Specialized resources are only used later, to improve certain notions.

The situation is the same for stylistic uses: [DUR 10] remarks that there are specific turns of phrase for each field. Certain syntactic constructions or collocations can be more frequent in a specialized discourse than in everyday language. It is often the case that the collocations specific to a field of specialization involve a different translation of one of the collocations. For example, répandre can be translated as to spread in everyday language; however, when talking about insecticide, we use the verb to spread in English but traiter in French. Unscheduled is translated in everyday French by imprévu but an unscheduled maintenance becomes un entretien curatif. The translation of prepositions can be delicate since a preposition can change the meaning of a term: exception detected by the program has a different meaning from exception detected in the programme. Moreover, the choice of preposition can be idiosyncratic for the term: we talk of filling out a form but this is achieved by filling it in. Once again [DUR 10] suggests that systematic research and documentation, which enables the translator to notice the linguistic uses specific to each field.

We will understand when reading [DUR 10] that a specialized translator spends part of his/her time researching documentation with the sole goal of manually creating terminological records, which will match not just the terms, but also their contexts (defining contexts for the meaning of the terms, “language” contexts highlighting the collocations and stylistic aspects).

[DUR 10]’s conclusions have been supported by other studies. Thus, [DAR 79] believes that a specialized language is specific not only by its naming convention, but also by what he refers to as his discourse:

When using what we call specialized languages, there is on the one hand the technical things that we have to be able to refer to exactly; and on the other hand, the text, which carries and actualizes these notions and

16 “On ne traduit pas une suite de mots, mais un message dont on a auparavant appréhendé le sens.”
which has to conform to certain form requirements. Therefore, the writer of the text has to have a dual competence: to be familiar with the naming convention of the topic and to be able to fully make use – in a certain register – of the language resources, which will highlight the elements of the naming convention. [...] With this in mind, we can consider that each specialized language presents itself under this dual aspect, and that naming convention and discourse must go hand in hand. Due to the appropriate documentation, it is often also easier to access the naming convention than the specialized discourse resources.

Echoing the distinction made by [DAR 79], [SCU 08] offers a fine analysis of the difficulties in translating French legal texts into Romanian. One point that we can highlight is that she considers that terms belonging to the naming conventions (e.g. technical terms) do not necessarily cause any difficulty in translation. [SCU 08] divides the naming convention into three categories:

**Words which are exclusively judicial terms:** these are technical terms used by the initiated. Some do not cause any difficulty in translation, because they have an immediate match in the target language (or are even borrowed from the source language) and have a formal resemblance to the source term, such as *abrogatif* and *abrogativ*. The terms that can be problematic in translation are terms that have no formal resemblance to the source term (e.g. *pronoce* and *pronunare*) and/or refer to a notion that does not exist in the culture of the target language (for example the French *communauté* and the Romanian *regim matrimonial legal*).

**Words with dual allegiances:** these are terms, which the legal system uses in very specific way. Among them we can mention:

- terms that are mainly judicial terms: these are judicial terms that have been appropriated by everyday language with a second meaning, such as *arbitrator, witness* and *guaranty*;

- terms that have a second legal meaning: these are terms whose main meaning is found in everyday language and who have acquired a specific meaning in the legal field, such as *act, mobile and enjoyment*.

The difficulty in translating these words with dual allegiances comes from the fact that they are shared with everyday language: their translation is only possible through context.

As for discourse, it covers several elements. We can find in it stylistic elements, specific phrasing and syntactic choices already highlighted by [DUR 10] as well as what [DAR 79] and [SCU 08] refer to as a support vocabulary. [DAR 79] defines support vocabulary as “the words whose technicality is low or non-existing and are used to actualize specialized words as well as provide the text with its organic
nature.” He lists the example of the legal field and the words “relationship breakdown”, “to hear a witness”, or “to cover” (in a financial sense).

Similarly, [SCU 08] mentions that – aside from words with dual allegiances – there are a certain number of terms that may not have a legal meaning but that, nonetheless, appear in texts with a specific meaning, which is different from the one they have in “common” language. For example, in French, affaire in legal documents does not have the meaning that is contained by its literal translation in Romanian (afacere). In a legal setting, it will be translated by cauz (so for example, “to take a case to court” would be in French porter une affaire devant la Cour and in Romanian a duce o cauz înaintea Curii, while “to do business” is translated by faire des affaires in French and a face afaceri in Romanian).

Let us point out that [SCU 08] and [DAR 79] both deplore the fact that the resources available to the translator do not take supporting vocabulary into account:

[We] can consider that each specialized language presents itself under this dual aspect, and that naming convention and discourse cannot be dissociated; it is often easier, due to the appropriate documentation, to access naming convention than specialized discourse resources. [DAR 79]

The books in question often only include the terms specific to the field itself and exclude terms of everyday language, which elude the neophyte’s understanding since they have acquired a specific meaning. [SCU 08]

According to [DAR 79], the lack of support vocabulary in technical glossaries can be explained by the fact that these resources are more aimed to help with understanding than with writing. Moreover, since the technical terms are striking due to their technicality, their presence naturally becomes necessary in a technical glossary. On the contrary, support vocabulary appears to be more transparent, and thus will be overlooked. However, it is just as essential. This point of view is also supported by [SCU 08]:

Paradoxically, when writing or translating a text it is often not the technical word, which is the biggest issue (these technical words have been and continue to be the focus of terminological lexicographies). We can notice by going through legal language indexes that many terms that were used when writing legal and administrative texts have not been retained. This is even truer if we take into consideration the situation of
bilingual dictionaries in the field. While it is true that in general the indexes target to understand more than writing. On the contrary, the terms of supporting vocabulary might appear to be marginal since they are transparent, but reveal themselves to be more delicate to manipulate, since they are required to go from mere lists of terms to a text: according to Darbelnet, *it is during the time of writing, and, therefore, we would add, during the time of translating, that this vocabulary effectively comes into its own* [DAR 79].

Since we situate ourselves in the perspective of facilitating translation and not knowledge engineering, our work does not focus on the extraction of translatable equivalences between terms. We focus instead on trying to identify the translations of any lexical element that might create translation difficulties. We will thus set aside in our research topic any information relating to syntax, style or text structure. We consider that any lexical unit that is not found in a general language dictionary might cause translation difficulties. Due to this definition, we exclude some terms that are frequently used in everyday language and whose translation has thus to be known by translators (e.g. *chemotherapy* is a medical term but its translation should not cause any issue for a professional translator). However we include elements such as *patient-centred*, which would not belong to a specific terminology but could create translation difficulties.

Therefore, from now on, our use of the word “term” should not be understood as corresponding to its official definition\(^\text{17}\) but as a “problematic unit for a technical translator”.

**1.2.5. Industrial context**

While there is a genuine qualitative interest in comparable corpora, they remain hard to use for translators. Compared to parallel corpora for which there are many existing tools, manual research and verification of informative contexts and translation matches in comparable corpora are time-consuming. This generates a loss in productivity and motivation for the translator.

There are very few computational tools that can help the translator when using comparable corpora. We can only mention two academic prototypes [BEN 00, SHA 06] and – as far as we know – when this doctoral research started, there was no commercial CAT tool that can process the comparable corpora. The technological transfer of the extraction techniques of bilingual lexicons from comparable corpora was our first task when we started to work as a research engineer

\(^{17}\) A term is a designation consisting of one or more words representing a general concept in a special language. [ISO 09].
for the LINGUA ET MACHINA corporation\textsuperscript{18}. This company was created by Emmanuel Planas based on his research results [PLA 98, PLA 00]; it edits the translation memory software SIMILIS [PLA 05] whose specificity resides in the fact that it uses linguistic analysis. The texts are labelled morpho-syntactically and the sentences are broken down into chunks. The matching of the segments of previously translated text and of text to be translated also happens at a linguistic level (matching the lemmas and grammatical categories) – and not at a graphical level, as is the case for other translation memory software.

LINGUA ET MACHINA also edits a web application for multilingual management in a corporate environment called LIBELLEX. This platform integrates several translation aid tools (bilingual concordance services, terminology extraction and management tools, translation memories, CAT and translation project management tools). The platform is different from SIMILIS, because it is designed to be not only for professional translators, but also for all the collaborators in the company (see Figure 1.1).

![Diagram of Libellex](image)

**Figure 1.1. Libellex: a multiservice platform for multilingual text management**

The possibility of using comparable corpora represents a major research and development axis at LINGUA ET MACHINA since fields of knowledge are constantly evolving and thus LINGUA ET MACHINA clients have to be able to quickly create translation resources even in fields in which there are little or no translation histories.

\textsuperscript{18} www.lingua-et-machina-com.
Part of the doctoral research work was thus to create a prototype which would enable the acquisition of bilingual lexicons from comparable corpora. We also developed an interface for looking up the extracted lexica. Generally speaking, the acquisition of bilingual lexicon from comparable corpora is carried out in two steps. First, the source and target terms are extracted from their respective corpora using terminology extraction techniques [BOU 94, DAI 94a, ENG 95]. Second, the terms extracted are aligned due to techniques based on the similarity of the terms’ occurrence contexts. We describe these techniques in section 1.3.

1.3. Term alignment from comparable corpora: a state-of-the-art

Specific approaches have been developed to acquire bilingual lexicons from comparable corpora. There are methods based on frequency distribution [KOE 02] or the use of semantic relations [JI 09] but we will not describe them here in detail, for either their results are not very convincing or they rely on advanced information extraction tools, which led us to decide that they would be difficult to implement.

Other methods look to extract parallel segments from comparable corpora [FUN 04, RAU 09]. While this approach may be efficient in creating general language lexica, it is hard to apply it to specialized corpora since it requires large corpora19. However, in a specialized field, the texts not only have to belong to the same field, but also be limited to a very specific topic, which means that gathering a large number of texts is almost impossible.

The most common state-of-the-art method to align terms in comparable corpora is called the distributional method or contextual similarity alignment method. We describe its principle in section 1.3.1.

1.3.1. Distributional approach principle

Distributional semantics, whose origins can be found in the works of Z. Harris, consider that it is possible to semantically characterize a word thanks to its distribution, i.e. all the words with which it fosters syntactic relations. Extraction of bilingual lexicons based on distributional semantics hypothesizes that it is possible to calculate the distributional similarities between words in different languages and that similar distributions correspond to semantic equivalences, no matter what languages are involved. This hypothesis was successfully tested by [RAP 95] and the first alignment model was presented by [FUN 97].

19 For example, [FUN 04] use a Chinese corpus of 110,000 sentences and an English corpus of 290,000 sentences to obtain 2,500 pairs of aligned sentences, with a precision index of 65.76%.
[RAP 95] shows the relevance of distributional semantics for the alignment of terms by demonstrating that there is a correlation between co-occurrence patterns of words observed in corpora from different languages: \(^{20}\) if words \(A\) and \(B\) co-occur in a significant manner in a language corpus \(L_1\), then their respective translations \(A'\) and \(B'\) in language \(L_2\) will also co-occur in a significant manner in a language corpus \(L_2\). For example, in a French–English medical corpus, we can expect *dépistage* and *radiographie* to co-occur in a significant manner in French, just like their respective translations – *screening* and *radiography* – do in English.

In his experiment, [RAP 95] represents the co-occurrences between the words by a matrix \(A_{ij}\) in which the value at the intersection of row \(i\) and column \(j\) refers to the normalized count\(^{21}\) of the co-occurrence of the word \(i\) with the word \(j\). The experiment he carries out starts with two such matrices. The first matrix contains the co-occurrences observed in the source corpus (English) and the other matrix contains the co-occurrences observed in the target corpus (German). At first, the two matrices are aligned, i.e. the word \(i\) of the English matrix is the translation of the word \(i\) of the German matrix. Then [RAP 95] randomly switches the order of the words in the matrices to misalign them. He then observes that the similarity\(^{22}\) of the source and target matrices decreases when the number of misaligned words increases.

[FUN 97] goes further with [RAP 95]’s experiment and uses a bilingual lexicon, which she projects onto the source and target corpora, which enables her to obtain attested translation pairs in both corpora. She then calculates a *context vector* for each source and target word whose translation is unknown. The context vector of a word \(w\) is an approximation of its distribution: it provides for each of the entries \(e\) of the bilingual dictionary a number of times in which \(m\) co-occurs with the entry \(e\) within a given contextual window (for example, three words to the right and three words to the left). Since the entries are attested in the source and target corpora, it is possible to compare the context vectors independently from their language. The closer two vectors are, the more plausible it is that their heads\(^{23}\) have a similar meaning and are translations of each other.

---

\(^{20}\) One should point out that the author, like many of the works mentioned in the following, did not directly rely on syntactic analyzers. On the one hand, syntactic contexts are only available if the corpus used for alignment was syntactically analyzed, which involves using costly development tools, which are rarely available. On the other hand, as soon as the corpora reach a sufficient size, the syntactic context can be approximated due to a window of \(n\) words surrounding the word to be semantically characterized.

\(^{21}\) The measure used is the mutual information described in Appendix A1.1.3.

\(^{22}\) The similarity between two matrices matches the sum of the differences between the values found in identical positions in the matrix.

\(^{23}\) Here we use the terminology given by [PRO 10] and will henceforth refer to the word whose distribution is calculated as the *head* of the vector and each entry of the bilingual lexicon presents an *element* of the context vector in the vector.
This alignment method can thus be summarized as follows:

1) Build the source and target term context vectors (see Figure 1.2):

   - the vector of a term $t$ matches $\tilde{t} = \{(m_1, \text{cooc}_1), \ldots (m_n, \text{cooc}_n)\}$ in which each $m_i$ is a co-occurring word with $t$ at the heart of a given contextual window (for example five words to the left and five words to the right of $t$) and $\text{cooc}_i$ is the number of times this co-occurrence happens.

2) Normalize the number of co-occurrences due to a measure such as mutual information or the likelihood ratio (see Appendix A1.1).

3) Translate the source term vectors in the target language due to a bilingual dictionary (see Figure 1.3).

4) For each source term (see Figure 1.4):

   - compare the translated context vector to the context vectors of the target words due to a similarity measure (see Appendix A1.2);

   - rank the target term vectors by descending similarity;

   - select the $N$ vectors, which are most similar: the target terms associated to these $N$ vectors are the candidate translations for the source term.

This alignment technique matches the state-of-the-art method, which was then used in a variation of ways, as we will see in section 1.3.3. However, before looking at these variations of the distributional method, let us first look at the evaluation methodologies of these alignment techniques.
1.3.2. Term alignment evaluation

The evaluation of alignment techniques is usually carried out by comparing the system outputs with a referential bilingual lexicon. The outputs of comparable corpora lexicon extraction system matches a list of pairs \((s, \{t_1, \ldots, t_n\})\) in which \(s\) is a source term and \(\{t_1, \ldots, t_n\}\) is the ranked set of its candidate translations. Contrary to the parallel corpora lexicon extraction systems, it is very difficult to obtain a good-quality lexicon simply by selecting the first candidate translation. The measures usually used for parallel corpora such as Alignment Error Rate [OCH 00] are not the most relevant since we try, here more specifically, to assess the algorithm’s ability to place the correct translation at the top of the list of candidate translations.
Literature presents three measures of evaluation: N-rank precision (also called \textit{TopN}), MRR (mean reciprocal rank) and MAP (mean average precision).

1.3.2.1. \textit{Precision at rank N or TopN}

This is by far the most common measure. It comes from the precision measure used in information retrieval. It represents the proportion of source terms, which have at least one correct translation in their \textit{N} first candidate translations:

\[
P_N = \frac{1}{|S|} \sum_{i=1}^{|S|} \alpha(T_{iN}, R_i) \tag{1.1}
\]

\[
\alpha(T_{iN}, R_i) = \begin{cases} 
1 & \text{if } T_{iN} \cap R_i \neq \emptyset \\
0 & \text{else}
\end{cases}
\]

in which:

- \textit{S} is the set of source terms with at least one candidate translation;
- \(T_{iN}\) is the set of the \textit{N} first candidate translations for the source term \textit{i};
- \(R_i\) is the set of reference translations for the source term \textit{i}.

It is also possible to calculate the recall on the \textit{N} first candidate translations \((R_N)\), which corresponds to the equation \([1.1]\), except that the set \textit{S} is the set of all the source terms, not just the source terms with at least one candidate translation. Increasing the precision makes the recall decrease, so \(F1_N\), which corresponds to the harmonic mean of \(P_N\) and \(R_N\), synthesizes the compromise between recall and precision [LAR 10b]. However, \(R_N\) and \(F1_N\) are not often used. The bilingual lexicon extraction systems are mostly term alignment systems: they take a set of source terms and target terms at the input and calculate a translation result for each source term, target term pair. In effect, at least one translation is suggested for each source term.

1.3.2.2. \textit{MRR}

The MRR measure matches the mean of the inverse of the correct translation ranks:

\[
MRR = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{\text{rank}_{i}} \tag{1.2}
\]

where \textit{S} is the set of source terms with less than one candidate translation and \(\text{rank}_{i}\) is the rank of the first correct candidate translation of the source term \textit{i} [YU 09].
1.3.2.3. **MAP**

[LAR 10b] suggest using MAP, which is also an information retrieval measure. It corresponds to the average of the precision value obtained for the set of top $k$ documents existing after each relevant document is retrieved, the average of the precision value obtained for the set of top documents existing after each relevant document is retrieved, and this value is then averaged over information needs:

$$\text{MAP} = \frac{1}{S} \sum_{i=1}^{S} \frac{1}{R_i} \sum_{j=1}^{R_i} \text{Precision}(T_{ij})$$

[1.3]

in which:

- $S$ is the set of source terms;
- $R_i$ is the number of referential translations for the term $S_j$;
- $T_{ij}$ is the set of candidate translations given by the system for the term $S_i$ with the reference translation $j$.

1.3.3. **Improvement and variants of the distributional approach**

Several variations and improvements on the distributional approach have been suggested. The latter focus on looking for distributional symmetry [CHI 04, SAD 03], the use of lexico-syntactic context (opposed to the contextual window) [OTE 05] and using anchor points, as is the case for parallel corpus sentence alignment [PRO 09]. Other variations have used second-order semantic affinities combined with semantic classes [DÉJ 02]. Finally, the work of [MOR 04] tried to align polylexical units.

1.3.3.1. **Favoring distributional symmetry**

[CHI 04] relies on the distributional symmetry hypothesis, which claims that “if two words are close in a translation direction as well as in the other (language A ↔ language B) then there are greater chances that they might be a translation of one another than if they are close only in one direction” [CHI 04].

[CHI 04] thus uses what she calls *crossed similarity* in opposition to *classic similarity* [CHI 04]. After carrying out two alignment processes, one process in the direction of source to target and the other process in the direction of target to source, [CHI 04] calculates, for each pair of source and target terms ($M_S$, $M_C$), the
harmonic mean of $rM_C$, $M_C$’s rank among the candidate translations of $M_S$ and $rM_S$, $M_S$’s rank among the candidate translations of $M_C$:

$$MH(rM_C,rM_S) = \frac{2 \times rM_C \times rM_S}{rM_C + rM_S}$$

[1.4]

Their experiments show that crossed similarity increases the number of translations found no matter what the type of the corpus (in the best case, the Top1 precision goes from 28% to 34%).

[SAD 03] also calculate a crossed similarity $SIM_{S\leftrightarrow C}$ between a source word $M_S$ and a target word $M_C$ based on:

$$SIM_{S\leftrightarrow C}(M_S,M_C) = SIM_{S\rightarrow C}(M_S,M_C) \times SIM_{C\rightarrow S}(M_C,M_S)$$

[1.5]

in which $SIM_{S\rightarrow C}(M_S,M_C)$ is the similarity between $M_S$’s vector translated into the target language and $M_C$’s vector, and $SIM_{C\rightarrow S}(M_C,M_S)$ is the similarity between $M_C$’s vector translated into the source language and $M_S$’s vector.

The authors also apply a morphological filter to translations: a noun can only be translated by another noun, a verb can only be translated by another verb, etc. Alignments thus obtained are assessed through their input in a cross-lingual information retrieval system: the use of the lexicon acquired by crossed similarity significantly increases the R-precision\textsuperscript{24} of the IR system by 27.1% (from 0.1417 to 0.1801) compared to the case in which the IR system only uses the lexicon acquired without crossed similarity.

1.3.3.2. Using syntactic contexts

[OTE 05] uses lexico-syntactic patterns acquired from a parallel corpus. For example, the pattern <import of [NOUN]> matches any noun appearing to the right of import of.

Bilingual syntactic patterns are acquired in three steps from an English–Spanish parallel corpus:

1) Acquisition of the English syntactic patterns on the source part of the corpus, for example: <import of [NOUN]>, <aid [VERB]>, <[NOUN] against fraud>

2) Acquisition of the Spanish syntactic patterns on the target part of the corpus, for example: <importación de [NOUN]>, <ayuda [VERB]>, <[NOUN] contra fraude>

3) Alignment of the English and Spanish patterns:

\textsuperscript{24} Average of precisions obtained for a recall level varying between 0 and 1.
<import of [NOUN]> → <importación de [NOUN]>
<aid [VERB]> → <ayuda [VERB]>
<[NOUN] against fraud> → <[NOUN] contra fraude>

The patterns are aligned using the Dice coefficient:

\[
\text{Dice}(\text{sourcepattern}, \text{targetpattern}) = \frac{2|S \cap C|}{|S| + |C|}
\]  

in which \( S \) corresponds to the number of sentences in which the source pattern appears, \( C \) corresponds to the number of sentences in which the target pattern appears and \( |S \cap C| \) corresponds to the number of times in which the source pattern and target pattern appear in the same aligned sentences. Only the pattern pairs with the best coefficient are retained (the threshold is determined empirically).

These bilingual patterns are used instead of a bilingual lexicon: the context vector of a word \( w \) contains, for each bilingual syntactic pattern \( p \), a score indicating how much \( w \) is encountered in \( p \). For example, if \( w \) is a noun, its context vector will indicate how often it is associated with the <import of [NOUN]> and <[NOUN] against fraud> patterns. The weight of the association between the head of the vector \( w \) and a syntactic pattern \( p \) is calculated from the number of times in which \( w \) instances \( p \),\(^{25}\) from the number of patterns instanced by \( w \) and the number of words instancing \( p \).

[OTE 05] obtains a precision of 89% on the Top1 and 96% on the Top5. These very good results can be explained by the nature of their data: the evaluation lexicon is made up of words whose number of occurrences is higher than 100 and the comparable corpus is composed of the unaligned parts belonging to a single-parallel corpus.

1.3.3.3. Relying on trusted elements

[PRO 09] uses anchor points, i.e. words used as trustworthy elements, for they are automatically identifiable, are not ambiguous and belong to the comparable corpus’ topic. The authors suggest giving them more weight than other elements in the context vectors due to their properties making them highly discriminating elements. Working from Japanese to French and English, they use transliterations and classical compounds as anchor points. The association measure between head and elements of the vector is the likelihood ratio. This measure is recalculated to favor anchor points. To this end, the sum of the likelihood ratio of a single vector is redistributed between

\(^{25}\)For example, in import of sugar the noun sugar instances the pattern <import of [NOUN]>. 
the co-occurrents in order to reinforce the anchor points and minimize those who are not:

\[
TV(M, m) = \begin{cases} 
TVI(M, m) + \beta & m \in PA \\
TVI(M, m) - \text{disparity}_M & m \notin PA
\end{cases}
\]

where \( TVI(M, m) \) is the initial likelihood ratio between the vector head \( M \) and its co-occurrent \( m \), \( PA \) is the set of anchor points, \( PA_M \) the co-occurrents of \( M \) which are anchor points, \( \neg PA_M \) the co-occurrents of \( M \) which are not anchor points, \( \beta \) is a coefficient varying between 1 and 20 (in the experiments, the best results were obtained with \( \beta = 8 \)).

Compared to the state-of-the-art method, the use of anchor points enables them to increase precision by +18% (from 17% to 20%) on the Top1 for English–Japanese translations and by +10% (from 20% to 22%) on the Top1 for French–Japanese translations.

1.3.3.4. Improving semantic information representation

[HAZ 12] notice that the way in which information is represented in the state-of-the-art approach is not optimal since it contains redundant and potentially incomplete information at the same time.

The authors suggest improving the context vector representation by applying a transformation by independent component analysis (ICA). This transformation enables them to generate a new-representation space in which the information is as independent as possible. The approach takes four steps:

1) Reduction of the matrix size by applying a principle component analysis (PCA).

2) Transformation by ICA of the matrix, taking into account the information of global nature (bilingual dictionary entry context) in order to obtain a representation space called GICA then calculation of the distance between source terms and target terms in this new space.

3) Transformation by ICA of the matrix, taking into account the information of local nature (target term contexts) so as to obtain a representation space called LICA then calculation of the distances between source terms and target terms in this new space.

4) Calculation of the distances between source and target terms by linearly combining the LICA and GICA distances.
The results obtained show that the GLICA approach provides better results than the state-of-the-art approach from Top6 when using the best parameter combination.\textsuperscript{26} The state-of-the-art approach obtains a precision level of 73.77\% on the Top20 and the GLICA approach obtains 75.40\% on the Top20. The approaches were tested on two corpora: a small specialized corpus and a large journalistic corpus.

1.3.3.5. Using second-order semantic affinities

The state-of-the-art method establishes the correspondence between the distribution of a source word and that of a target word by direct translation: each of the co-occurrences of the source word is “mapped” into the target corpus through the bilingual lexicon, then one tries to find a target word with a similar distribution to that of the “mapping”. This method is highly dependent on the bilingual lexicon’s coverage: only elements present in the lexicon and the two corpora will be present in the context vectors.

To overcome this issue of coverage, [DÉJ 02] suggests a method using distributional similarities between the terms to be aligned and the bilingual lexicon entries. They consider two terms whose semantic proximity with the bilingual lexicon entries are similar to be potential translations. The alignment method can be broken down into five steps:

1) Build the context vectors for the source terms and target terms to be aligned.

2) Build the context vectors for the source words and target words present in the bilingual dictionary.

3) Build, for each source term, and target term, its similarity vector: this vector indicates – for each of the entries \( e \) in the bilingual lexicon – the similarity between the context vector of the term and the context vector of the entry \( e \). The size of the similarity vector can be parametrized, i.e. we can choose to only retain the \( n \) entries the most similar.

4) For each pair \((\text{source term}, \text{target term})\), calculate the similarity between their respective similarity vectors.

5) For each source term, select the \( N \) target terms whose similarity vector is the closest to the source term’s similarity vector.

This method thus enables the authors to translate any word in the corpus, even if no other element of the vector can be translated.

\textsuperscript{26}For the state-of-the-art approach, the best results are obtained with the likelihood ratio as the normalization measure of the co-occurrences count and the Jaccard as similarity measure. For the GLICA method, the best results are obtained with mutual information as the similarity measure. The similarity measure is the standardized Euclidian distance. These measures are given in Appendix 1.
[DÉJ 02] call their method “interlingual similarity translation”\textsuperscript{27} and oppose it to the “direct translation”\textsuperscript{28} advocated by [FUN 97]. Indeed, what is projected in the target language is the level of similarity between the lexical entries and the word to be translated, and not the context of the word to be translated.

The results obtained are not necessarily convincing. At best, the interlingual method enables the authors to obtain 51 correct translations on the Top20 whereas the state-of-the-art method obtains 57.

1.3.3.6. Improving the bilingual resource with semantic classes

[DÉJ 02] experiment with the use of the semantic classes of a thesaurus in combination with the interlingual method (see section 1.3.3.5). Instead of using a traditional bilingual resource, the authors use a thesaurus. The thesaurus is used to include new entries in the similarity vectors used in the interlingual method. For a term to be aligned \( t \) matching the similarity vector \( vs \), the inclusion of new entries happens as follows:

1) \( E \) is an empty starting set.

2) Select the \( n \) entries of the thesaurus the closest to \( t \), these entries are the set \( E_0 \).

3) For all the entry pairs \( (e_1, e_2) \) in \( E_0 \):

   - add to \( E \) all the thesaurus entries, which can be found on the minimal path between \( e_1 \) and \( e_2 \);
   - add \( e_1 \) and \( e_2 \) to \( E \).

4) Add all the entries found in \( E \) to \( vs \).

In the end, this technique enables them to obtain 63 correct translations on the Top20 against 57 for the state-of-the-art method.

1.3.3.7. Translating polylexical units

The works mentioned until now only focus on the translation of monolexical units (i.e. units made of a single word). [MOR 04] suggest adapting the interlingual similarity approach to polylexical units. This method is interesting for us, since we will need to translate such units within the context of CAT.

[MOR 04] suggest building the polylexical unit context vector as the union of the context vectors of each word which makes up the polylexical unit. The alignment method used also calls upon second-order semantic affinities:

1) Build the context vectors for source monolexical units.

\textsuperscript{27} “traduction par similarité interlingue”.

\textsuperscript{28} “traduction directe”.
2) Build the context vectors for source polylexical units as the union of the context vectors of each of the words that makes them up.

3) Build context vectors for target monolexical units.

4) Build context vectors for target polylexical units as the union of the context vectors of each of the words that makes them up.

5) For each source unit to be translated (see Figure 1.5):
   - select the $n$ entries in the bilingual lexicon, which are the closest to the source unit;
   - select the target context vectors in the $n$ bilingual entries;
   - calculate the barycenter of these $n$ target vectors: we obtain a mean context vector in the target language;
   - compare this mean vector to the context vectors in the target units;
   - select the most similar of the $N$ vectors: the heads of these $N$ vectors are the candidate translations for the source unit.

![Figure 1.5](image)

This method provides good results for the polylexical terms whose translation is also a polylexical term (88% on the Top20). The results are more limited for polylexical terms whose translation is either a monolexical or a polylexical term (55% on the Top20). By comparison, the precision level obtained for monolexical terms whose translation is also a monolexical term is 51% on the Top20.
<table>
<thead>
<tr>
<th>REFERENCE</th>
<th>LANGUAGES</th>
<th>ELEMENTS TO BE TRANSLATED</th>
<th>SUBJECT DOMAIN</th>
<th>BILINGUAL RESOURCES</th>
<th>METHOD</th>
<th>PRECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>[RAP99]</td>
<td>DE → EN</td>
<td>100 UML</td>
<td>press</td>
<td>298M</td>
<td>direct</td>
<td>.65 .89</td>
</tr>
<tr>
<td>[CHI02]</td>
<td>FR → EN</td>
<td>95 UML</td>
<td>≥ 100 medical SC</td>
<td>1.2M</td>
<td>direct</td>
<td>.13 .61 .94</td>
</tr>
<tr>
<td>[DÉJ02]</td>
<td>DE → EN</td>
<td>1800 UML</td>
<td>medical SC</td>
<td>200k</td>
<td>direct</td>
<td>.44 .57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>180 UML high</td>
<td>social sc. SC</td>
<td>8M</td>
<td>interlingual</td>
<td>.43 .51</td>
</tr>
<tr>
<td>[MOR04]</td>
<td>FR → EN</td>
<td>100 UML</td>
<td>≥ 5 environment</td>
<td>4.9 M</td>
<td>interlingual</td>
<td>.41 .51 .45 .55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 UPL translated by UML or UPL</td>
<td>≥ 5 TECH</td>
<td>22 k LG</td>
<td>.87 .88</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 UPL translated by</td>
<td>≥ 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[MOR07]</td>
<td>FR → JA</td>
<td>100 UML</td>
<td>≥ 2 medical SC + VG</td>
<td>1.5M</td>
<td>.51 .6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 UML et UPL</td>
<td>≥ 2 medical SC</td>
<td>659k</td>
<td>.3 .42</td>
<td></td>
</tr>
<tr>
<td>[PRO10]</td>
<td>FR → EN</td>
<td>122 UML</td>
<td>≥ 5 medical SC</td>
<td>507k</td>
<td>.21 .47 .57</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>648 UML</td>
<td>≥ 15</td>
<td>173k LG, medical</td>
<td>direct</td>
<td>.13 .34 .41</td>
</tr>
<tr>
<td>[HAZ12]</td>
<td>FR → EN</td>
<td>122 UML</td>
<td>≥ 5 medical SC</td>
<td>530k</td>
<td>.34 .64 .76</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>500 UML</td>
<td>press</td>
<td>244k LG</td>
<td>ICA</td>
<td>.16 .32 .40</td>
</tr>
</tbody>
</table>

**Table 1.1. Results of the state of the art - alignment by contextual similarity**
We have seen in this section different alignment techniques – the state-of-the-art method, the interlingual method – as well as several variations on these methods – use of semantic classes, of anchor points, use of lexico-syntactic contexts, etc. Table 1.1 offers a synthesis of the results of this work. This table also specifies the type of terms to be translated (monolexical, coded as UML or polylexical, coded as UPL) as well as the size and nature of the corpora used: generalist texts (LG), scientific (SC), popular science (VG) or technical (TECH) specialized texts.

We can see that beyond the alignment techniques, the data used as well as the various parameters also influence the quality of the results. This impact is analyzed in section 1.3.4.

1.3.4. Influence of data and parameters on alignment quality

[LAR 10b] already provided a very good overview of the impact of data and parameters on the quality of the extracted lexicon. We have here added to their observations with an analysis of the results obtained through the approaches described in section 1.3.3. We start by describing the impact of the data (section 1.3.4.1) then the impact of the parameters (section 1.3.4.2).

1.3.4.1. Data

The factors which influence the quality of alignments the most are linked to the nature of the data:

Frequency of the elements to be translated – an element is all the better translated when it is frequent: its context is calculated from a greater number of occurrences and thus it is more representative and gives a better semantic characterization of the element to be translated. This is particularly well shown in the experiments by [PRO 10]: the less frequent words (maximum 25 occurrences) obtain approximately 7% on the Top20 whereas the most frequent occurrences (beyond 800 occurrences) obtain a score of 100% on the Top20.

Specialization of the elements to be translated – [CHI 04] mentions that the specialized elements are better translated than the general language elements, no matter their frequency. A similar result is obtained by [HAZ 12]: the alignment obtained are of better quality for the specialized corpus even though the press corpus is much bigger. This can be explained by the fact that the terms are generally semantically straightforward whereas polysemy or meaning variations are frequent in common words: this means that the context vector is more “fuzzy” and less discriminating.

Size of the corpus – when comparable corpora are large, the terms to be translated generally occur a great many times. This enabled us to build more representative
context vectors. But the size in itself is insufficient. Corpora must also be sufficiently comparable.

*Comparability of the corpora* – [LI 10] defined a corpus comparability measure. This measure shows the expectation of finding the translation of a source word in the target corpus (and vice versa). It is based on the projection of a bilingual dictionary in the corpus (see Appendix A1.3).

Relying on this measure, [LI 10] shows the impact of the comparability of corpora on the precision of alignments. [LI 10] starts from an original corpus called $C$ from which they extract two highly comparable corpora called $C^1$ (comparability of 0.882) and $C^2$ (comparability of 0.916). The lexicons extracted from $C^1$ and $C^2$ are of better quality than those extracted from $C$: [LI 10] increase the precision of the results by between 5.3% and 9.5% on the Top20.

*Specialization of the bilingual lexicon* – in their experiments, [LAR 10b] compared the results obtained depending on the degree of specialization of the entries in the bilingual lexicon used to translate context vectors. For a bilingual lexicon of 5,000 entries, the results are slightly better when the lexicon is partially made of specialized lexies than when it is only composed of entries belonging to the general language (F1-measure on the Top1 goes from 38.9 to 39.4 and MAP goes from 0.471 to 0.473). It is the same for [PRO 10] who decides to reinforce the specialized elements (anchor points). In a personal experiment, we also observed that the presence of specialized elements in the bilingual lexicon improved the results (see section 1.4.1.5).

[CHI 04] obtains contradictory results. She mentions that adding words from the general language to the bilingual lexicon improves the results (from 59.4% to 100% on the Top20). This is particularly visible when the elements to be translated are terms and not words from the general language. However, the first lexicon is made of 4,963 entries, which only belong to the medical field, and the “improved” lexicon contains 6,210 entries belonging to the medical field and the general language. It is hard to say if the improvement of the results is only due to the addition of general vocabulary in the bilingual lexicon or the simple increase in the number of entries.

1.3.4.2. *Parameters*

The parameters involved in the distributional approach are:

– size and nature of the context in which the co-occurrences are gathered: sentence, paragraph, window of $n$ words around the vector’s head and syntactic context;

---

29 For the comparable corpus $C$, if we consider the translation process from the English part $C_e$ to the French part $C_f$, a comparability measure $M_{e,f}$ can be defined on the basis of the expectation of finding, for each English word $w_e$ in the vocabulary $C_e^v$ of $C_e$, its translation in the vocabulary $C_f^v$ of $C_f$. 645.
– normalizing the co-occurrences count: several measures are possible such as the likelihood ratio, the mutual information or the TF-IDF. These measures are detailed in Appendix A1.1;

– calculating the similarity between vectors, for example: cosine measure, Jaccard and Euclidian distance. These measures are described in Appendix A1.2.

These parameters are complicated to manipulate. [PRO 10] shows that the optimal combination of parameters depends on the corpus used and the languages involved and that it is impossible to determine it in advance.

Moreover, [LAR 10b] mentions that the choice of context can also depend on the final use of the bilingual lexicons. In their experiment, if the context matches the paragraph, they obtain a very good recall on the Top20, which can be useful for the semi-supervised creation of linguistic resources. However, if the context matches the sentence, we obtain a higher precision, which is ideal for the unsupervised creation of bilingual lexica.

[HAZ 12] shows that in a press corpus of large size, the best results are obtained with the likelihood ratio whereas on a small specialized corpus, the best results are obtained with mutual information. Only the size of the context can be anticipated: the elements with low frequency are better translated when the context vector is calculated in a short window [PRO 10].

A solution to this parameter problem would be to learn the best parameter configuration before extracting the lexicon. Several configurations could be tested by using translation pairs present in the bilingual lexicon as evaluation lexicon. The best configuration thus obtained could then be applied to align the terms.

1.3.5. Limits of the distributional approach

While parallel corpus alignment tools generate accurate translation pairs in more than 80% of the cases [DAI 94b, MAC 08, VIN 10], this is far from the case for comparable corpus alignment. We have seen in the previous section that the results obtained with comparable corpora vary between 30% and 89% on the Top10 and between 40% and 94% on the Top20 depending on the language pairs, the volume and quality of the data, the nature and frequency of the elements to be translated.

This contrast between the results obtained in comparable corpora and the results obtained in parallel corpora can be explained for two reasons:

Research space – in a parallel corpus, the research space is progressively diminished: one starts by looking for anchor points (cognates, figures) then aligns sentences, and then the translations of the terms are looked for within pairs of aligned
sentences. In a comparable corpus, the translations of the terms are looked for in all the corpus.

Translation presence – in a parallel corpus, unless the translator has forgotten to translate a word, the source term always has a translation. In a comparable corpus, not only can a term not have any translation, it is also very hard to determine whether or not this translation might be present.

In addition, there are limitations inherent to the distributional method:

Semantic homogeneity of the vectors – if the element to be translated is polysemic or presents variations of meaning, its context vector will be less semantically homogeneous, since the element is used in varied contexts.

Term frequency – the element to be translated and its translation must be sufficiently frequent: the more the vectors are built from a great number of co-occurrences, the more representative they are of the distribution of the term.

Things are even more complicated when the alignment happens in a specialized corpus to acquire bilingual terminology lists:

Prior terminology extraction – there is a reliance on the term extractor: the target term can be found in the target corpus without having been extracted by the extractor.

Relevance of the bilingual dictionary – the bilingual dictionary used for the transfer can contain translations which are not appropriate in the subject domain of the texts and contribute to bias the projection of the source vector into the target language.

Size of the corpus – specialized corpora, since they correspond to a well-defined topic, are often small: their volume is closer to the hundreds of thousands of words [PRO 10] than the millions of words [RAP 99]. Therefore, the terms occur less often and their vectors are less representative.

Polylexical terms – we also try to align polylexical units. However, as it is mentioned in [MOR 07], complex terms have lower frequencies than simple terms, which make their context vectors less representative. If the complex term vector is a vector made of the vectors of the lexical words that create the term, then this lowers the semantic homogeneity of the vector.

In this section, we have presented a state-of-the-art of the comparable corpus alignment techniques. In section 1.4, we will describe the way in which we have created a CAT prototype that relies on the distributional method to extract bilingual lexicons from comparable corpora.
1.4. CAT software prototype for comparable corpora processing

In the industrial context of LINGUA ET MACHINA, comparable-corpus extraction is meant to provide impetus for the generation of linguistic resources in emerging fields or fields in which the corporation has very little translation memory. The provided corpora should be small specialized corpora (less than 2 million words). The precision scores thus obtained would be at best between 34% on the Top1 and 76% on the Top20. We can immediately anticipate the fact that translators will not be satisfied with a simple list of source terms and candidate translation alignments. To overcome these uncertain results, it is necessary to accompany these alignments with various pieces of information presented as a terminology record, which will allow the translation to decide which candidate translation is the right one.

The developed prototype is shown in Figure 1.9. It is able to extract terms from texts in the source and target languages and align them with a method based on the distributional approach (section 1.4.1). Then the prototype collects information from the texts in the corpus and on the web, which will be offered to the translator as a terminology record (section 1.4.3). A user interface for looking up the extracted lexicons was also developed (section 1.4.2).

1.4.1. Implementation of a term alignment method

1.4.1.1. Implementation and data

The point of this work is to create a first simple prototype, which could then be used to observe how the translators approach the lexicons extracted from comparable corpora and assess the contribution of these lexicons to specialized translation (Chapter 2). We have chosen to implement the state-of-the-art method due to its ease of implementation. The series of variations suggested required either specific resources (parallel corpora to learn lexico-syntactic patterns, thesauri), or the development of pre-processing tools (transliteration and neoclassical compound extraction), or these methods are time-consuming, which is always an issue in an industrial environment (crossed similarity, interlingual similarity, GLICA method). From all of the approaches mentioned, we will here only keep [MOR 04]'s approach for the alignment of polylexical terms. As it was demonstrated by [PRO 10], at the current level of research, it is impossible to decide at first what will be the best combination of parameters to use. In a corporate environment, we always apply the same parameter, regardless of the size or topic of the corpora. We have arbitrarily chosen to use the Jaccard index as the similarity measure and the likelihood ratio for normalizing the co-occurrences count. However, we have carried out a few experiments regarding the interactions between the size of the contextual window and the frequency of the terms to be translated. We have also assessed the input of specialized resources, which can be a simple and efficient means to improve the quality of lexica.
These tests were carried out on a small specialized English–French corpus (approximately 400,000 words per language) concerning breast cancer. This corpus is described in more detail in sections 2.3.2.1.1 and 5.2. To validate our experiments, we used two reference lexica:

Specialized lexicon: 177 English–French pairs of monolexical terms collected by [PRO 10] from the UMLS\textsuperscript{30} and the \textit{Grand dictionnaire terminologique.}\textsuperscript{31}

Generalist lexicon: 1,842 English–French pairs extracted from our bilingual dictionary.

45 entries overlap between the two lexica. In accordance with the other research works, we ensured that each term to be translated appeared at least 5 times in the corpus. Translation was carried out from English to French.

The bilingual lexicons acquisition method implemented can be broken down into four steps:

– Extraction of the terms to be aligned (section 1.4.1.2).

– Gathering the context vectors (section 1.4.1.3).

– Translating the context vectors (section 1.4.1.5).

– Aligning the terms (section 1.4.1.6).

1.4.1.2. Extraction of the terms to be aligned

The terms to be aligned are extracted from source and target corpora. These terms are either polylexical terms extracted by the terminology extractor\textsuperscript{32} integrated within SIMILIS (LINGUA ET MACHINA’s translation memory software), or monolexical terms (e.g. simple words) belonging to the grammatical categories of noun, adjective, adverb and verb, occurring more than five times. The minimum threshold of five times was chosen for two reasons: (1) it reduces the number of terms to be aligned and thus the processing time; (2) it is the minimum occurrence number chosen in the research works presented in section 1.3. We believe that below five occurrences, the context vector is not significative. The pre-processing of the corpus (tokenization, lemmatization and part-of-speech tagging) is carried out by the linguistic analyzer XELDA.\textsuperscript{33}


\textsuperscript{31} www.granddictionnaire.com.

\textsuperscript{32} While this tool is called terminology extractor by LINGUA ET MACHINA, it does not actually extract terms \textit{per se} (e.g. referring to a field-specific concept) but rather noun and verb phrases.

\textsuperscript{33} www.temis.com.
1.4.1.3. Collecting context vectors

1.4.1.3.1. Monolexical term context vectors

The size of the context was chosen after several attempts on our corpus. Contrary to [PRO 10], we have not observed any influence of the frequency of the terms to be translated on the size of the ideal contextual window (see Figure 1.6). In our prototype, the size of the context will thus be of three lexical words to the left and three lexical words to the right of the term to be translated, regardless of its frequency.

![Graph showing influence of window size on precision](image)

The units to be translated match the terminology entries and generalist entries. Each curve matches a frequency range, and the number of entries in this frequency range is mentioned in brackets.

**Figure 1.6. Influence of the frequency of terms to be translated on the size of the optimal contextual window**

The number of co-occurrences is standardized with the likelihood ratio [DUN 93]. Its calculation is detailed in Appendix A1.1.1.

1.4.1.4. Polylexical term context vectors

We calculate the context vector of a polylexical term, just like [MOR 04] did, from the context of each lexical term that is contained in it. [MOR 04] calculates the union of these vectors; as for us, we calculate an average vector:34

- The term *breast cancer* has two lexical words: *cancer* and *breast*
- Their context vectors are:
  - cancer = \{(*cancer \leftrightarrow 50), (*breast \leftrightarrow 30), (*treatment \leftrightarrow 25)\}
  - breast = \{(*breast \leftrightarrow 60), (*cancer \leftrightarrow 30), (*ablation \leftrightarrow 20)\}\n
34 This leads to a gain of about two points of precision on the Top20, but we do not know if this is specific to our corpus or independent of the data.
– The context vectors of breast cancer is thus:

\[
\text{breast cancer} = \{(\text{breast} \leftrightarrow 45), (\text{cancer} \leftrightarrow 40), (\text{treatment} \leftrightarrow 12.5), (\text{ablation} \leftrightarrow 10)\}
\]

1.4.1.5. Translation of the source context vectors

Source context vectors are translated into target languages with several dictionaries. First, we use the bilingual dictionary integrated in our linguistic analyzer (XELDA). This dictionary has 37,655 entries (an English word is translated by 1.58 French words on an average). The size of the dictionary was increased with translation links extracted from Wikipedia and Wiktionary. This enables us to translate 18% more words in context vectors. Figure 1.7 shows the input of these two resources.

![Graph showing the influence of the bilingual dictionary](image)

**Figure 1.7. Influence of the bilingual dictionary (Specialized lexicon)**

When a context word has several possible translations, its likelihood ratio is distributed over all its translations according to the frequency of each translation in the target corpus. For example:

– the vector of patient will contain the association (related \(\leftrightarrow\) 60);

– related can be translated by:

  - parent: 10 occurrences in the target corpus,
  - proche (kin): five occurrences in the target corpus,

– The vector of patient thus translated into the target language will contain the associations:

  - (parent \(\leftrightarrow\) 40),
- \( (\text{proche} \leftrightarrow 20) \).

When several context words match one translation, the likelihood ratios add up. For example:

- the vector of \textit{patient} contains the association \( (\text{rebuilding} \leftrightarrow 10) \) and the association \( (\text{reconstruction} \leftrightarrow 20) \);
  - \textit{rebuilding} is translated by \textit{reconstruction},
  - \textit{reconstruction} is also translated by \textit{reconstruction},

- the \textit{patient} vector translated in the target language will thus contain the association:
  - \( (\text{reconstruction} \leftrightarrow 30) \).

1.4.1.6. Term alignment

The similarity measure used to compare the source and target vectors is the Jaccard index [MOR 04], which is detailed in the Appendix A1.2.2. The alignment is carried out from English to French and we retain the Top 20 best candidate translations.

[MOR 04]’s experiments on the alignment between monolexical and polylexical terms had produced limited results, so we have separated these two types of unit. The monolexical terms can only be aligned with other monolexical terms and the polylexical terms can only be aligned with other polylexical terms.

For the monolexical terms, we added a grammatical category filter (as [SAD 03] had): nouns can only be aligned with nouns, adjectives with other adjectives, etc.

Figure 1.8 shows the differences in the results obtained on these two lexica. We can see that the terms from the specialized lexicons are better translated than those of the generalist lexica. Two reasons can explain this: (1) specialized entries are on average more frequent (166 occurrences against 54 for generalist entries), which means their context vector represents them better; (2) as specialized vocabulary, these terms are probably less subject to polysemy, which means their context vectors are more homogeneous.

1.4.2. Terminological records extraction

As we mentioned in section 1.2.3, a simple list of alignments is not sufficient for translators: they need to be able to access information, which will recontextualize the term and allow them to understand its meaning.
This is the reason why we developed a terminological record extraction module, which gathers the following information for each term:

*The entry term* – this is the lemma for monolexical terms and the most common inflected form for polylexical terms.

*Part-of-speech* – we display the part of the speech tagged by XELDA (monolexical terms) or a grammatical type tagged by SIMILIS (polylexical terms).

*Frequency* – the number of occurrences or frequency (number of occurrences divided by the number of words in the corpus) does not mean much to translators. We chose instead to use the three classes calculated from the distribution of the occurrences of lexical words:

– frequent use (number of occurrences of the term is above the 90th percentile);
– infrequent use (the number of occurrences ranges from the 51st percentile to the 90th percentile);
– rare use (the number of occurrences is below or equal to the 50th percentile).

*Definition* – when it exists, we provide a link to the Wikipedia or Wiktionary page.

*Collocations* – collocations are words whose appearance in the left or right context of the term is significantly frequent (we used the likelihood ratio to compute this). For example, lymph node is linked to axillary lymph node. To select collocations, we rank all the collocations found in the corpus by descending likelihood ratio and only keep those that belong in the top 25%.

*Context* – these are all the sentences in which the term appears. The term is set in bold in the sentence. A link to the document from which the sentence is taken is provided. Contexts are not ranked.
Variations – these are simply spelling variations (e.g. hyphens, alternation between -or / -our in English).

Close terms – these are the terms, which have at least one lexical word in common with the key term, for example tumor is linked to benign tumor, tumor growth, etc.

1.4.3. Lexicon consultation interface

Translators can look up the lexicons in a dedicated interface. Screen captures are provided in Appendix 3 and the prototype can be freely seen at the following address: http://80.82.238.151/Metrict/InterfaceValidation/.\(^{35}\) This consultation interface offers tools to facilitate term searches. The translator has a query field, which allows him/her to explore the lexicon at ease, and “fuzzy” queries are possible. For example, the query “lymph%” will bring up all the terms starting with lymph-. If none of the terms match the request, a search is carried out directly in the corpus’s texts.

![Diagram of the consultation interface](image)

**Figure 1.9. Implementing a method to acquire bilingual lexicons and a search tool of the extracted lexica**

1.5. Summary

This chapter started with a short historical overview of the beginning of MT. We saw that the field of the translation aid – a field in which we have carried out our

---

35 The user name is “test”. Leave the password field empty.
research – arises from the first disappointing machine translation tools. Computer-aided translation tools have a less ambitious but more realistic goal: the objective is to provide the translator with a tool to improve his productivity, not to replace him/her.

Until recently, translation aid software always required the existence of a translation memory to operate correctly. This created issues when there was no such translation memory (languages with few available resources, emerging fields). Moreover, research in translation studies has shown that – due to quality reasons – translators prefer to access multilingual corpora of texts, which have not been generated by a translation (comparable corpora). These comparable corpora allowed them to gather information about linguistic uses and terminology used in the technical field on which they were working. In spite of this obvious need, there are actually very few tools that can help the translator to explore comparable corpora and identify translation pairs.

The first part of the research carried out here was to create a prototype of such a tool. After reflecting on the state-of-the-art of alignment techniques based on comparable corpora, we have opted for a state-of-the-art method. This method does not require any pre-processing or specific linguistic resources and is more time-efficient than other more elaborate methods.

Once the terms are aligned, our prototype automatically generates terminological records, which have various pieces of information, which might help the translator. A consultation interface was developed to facilitate resource access.

While translation studies confirmed the appeal of comparable corpora to check translation hypotheses or better approach a technical field, we do not know how these corpora and their corresponding lexicons can be used as the only translation assistance tool (case in which there is no translation memory). This is the direction in which we take our research in Chapter 2.