Hartwell, L. M. & O. Kraif. (2022). Parallel corpora in English language teaching. In R. R. Jablonkai et E. Csomay (Dirs.) *The Routledge handbook of corpora in English language teaching and learning (*p. 478-494). London: Routeledge.

Laura M. Hartwell, ORCID: 0000-0002-8601-4176, Lairdil, Université Toulouse Capitole

Olivier Kraif, ORCID: 0000-0002-8935-7342, Lidilem, Université Grenoble Alpes

Author's last draft

1. Introduction

Since the Rosetta stone, multilingual texts have been essential for understanding other languages, notably bilingual editions for ancient Greek and Latin, as well as for European literature, such as the French *Les Belles Lettres* or *l'Aubier* (Chartier & Martin, 1986). "Dual language" or "bilingual" books are used positively for foreign or second language learning by children and adults in and outside of the classroom (Ernst-Slavit & Mulhern, 2003). Instead of reading full texts in more than one language displayed side-by-side as for bilingual books, a parallel corpus query allows users to consult multiple occurrences of a keyword or sequence and their translations, relying upon statistical measures to identify the likelihood of those constructions.

Much of parallel corpora research stems from human and machine translation studies. In her evaluation of the usefulness and usability of parallel corpora for research and translation needs, Rabadán (2019) offers "action points" also relative to teaching and learning: determining clear questions, reviewing available corpora and their reliability, building upon existing resources, being creative, contributing and sharing resources (p. 71). Translation studies and teaching draw upon parallel corpora to determine how best to render an enunciation from a source to a target language. Language learning requires understanding and expressing one's self in a target language. The border between translation and teaching is blurred as professionals and students increasingly rely upon available automatic translators.

However, expressing one's self extends beyond translation equivalents, requiring attention to both audience and cultural differences. Thus, parallel corpora invite new means to examine how language works across languages and across cultures.

A parallel corpus is composed of source texts aligned with their translations by word, chunk, sentence or paragraph in one or more other languages. The order of the languages in the translation process may be either unidirectional (i.e. from one language to another), bidirectional or "reciprocal" (i.e. translations from and to both languages), multidirectional (i.e. the source and/or translations including a range of languages) or translation via a third language (for example, a source document in Swahili, translated to English, before translation to Italian). Translations have typically come from available fictional works, governmental records, business needs, and film subtitles. This availability reflects societal advancements and has largely influenced both the creation of parallel corpora and the possibilities for research (Doval & Sánchez Nieto, 2019). Essential to translation and translation studies, parallel corpora are increasingly consulted directly or in mediated forms for foreign and second language teaching and learning (Frankenberg-Garcia, 2012a).

In this chapter, we review current research that relies considerably on available existing sources, notably institutional productions and translation studies. We discuss how the availability of tools and languages define factors of core issues. Based on these factors, we focus on language teaching and learning, including three case studies illustrating the interest of parallel corpora. We conclude with perspectives and recommendations. First, we highlight the differences between parallel corpora and similar resources.

1.1 Difference from comparable multilingual corpora and dictionaries

A parallel corpus differs from a comparable multilingual corpus. A comparable corpus is a collection of independent texts in two or more languages, containing no translations from the

other language. For example, the corpus Étude interdisciplinaire et interlinguistique du discours académique (EIIDA) (Carter-Thomas & Jacques, 2017) is a comparable corpus of research articles and conference transcripts in either English or in French. All of the articles and transcripts in the EIIDA corpus are in their original language, allowing comparisons, such as word frequency, pronoun use or overall structure across languages. A comparable corpus contains no translations of the original documents, thus making automatic alignment at the word or sentence level impossible. This automatic alignment is a key technical element of parallel corpora as it allows users to scrutinize languages at the word or sentence level.

Within lexicology, Teubert and Čermáková (2004) explain how parallel corpora differ from multilingual dictionaries by their larger variety of translation equivalents that reduce ambiguity:

> Even the largest bilingual dictionary will present only a tiny segment of the translation equivalents we find in a not too small parallel corpus. Because the ordering principle of printed dictionaries is alphabetical, based on mostly single-word entries, bilingual dictionaries do not record larger and more complex units of meaning in a methodical way. Neither do they tell us which of the equivalents they offer belong in which contexts. […] From parallel corpora, we can extract a larger variety of translation equivalents embedded in the contexts, which make them unambiguous. This is what makes parallel corpora so attractive. (p. 123).

1.2 An overview of technical considerations

Parallel corpora share commonalities with unilingual corpora in their reliance upon technical advances. The early 1970s marked the first multilingual corpus, the Yugoslav-Serbo-Croatian-English Contrastive corpus compiled by Filipovac (Doval & Sánchez Nieto, 2019). However, the birth of parallel corpora was the English-Norwegian Parallel corpus, of which

the English-Swedish Parallel Corpus built upon the same design and many of the same English documents (ibid., p. 2). The original corpus was compiled between 1994 and 1997, later tagged for parts of speech and enlarged with texts in German, Dutch and Portuguese, especially extracts from novels and non-fictional books, as part of the password-protected Oslo Multilingual Corpus (Johansson, 2008).

Creating parallel corpora remains more technically complex than creating unilingual corpora. Beyond the issue of gathering quality bilingual sources, issues of alignment (i.e. by word, chunk, sentence or page) to the corresponding translation raises further issues, as we will discuss, since translations are not simple word-to-word correspondences and have possibly different alphabetic or character systems.

2. Review of current sources inspiring research

The field of parallel corpora research is young and inextricable from the creation of consequential parallel corpora. Researchers rely upon sufficiently large authentic bilingual or multilingual resources. The compilation of these corpora precedes and determines the current research, as discussed next.

2.1 Governmental and intergovernmental sources

Governing bodies publish high quality documents in many languages allowing their community to access key legal and other documents. Thus, the United Nations' official records are available as parallel corpora in Arabic, Chinese, English, French, Russian and Spanish. These and other documents, such as those of the European Parliament, are valuable downloadable multilingual data sources. They are also incorporated within research-oriented platforms.

One early source is the Canadian government's collection of laws and other governmental documents in both English and French since 1978, including 'Hansards' (official records). These Hansards were incorporated into initial systems of bilingual aligning, notably within the Bitextes Anglais-Français corpus, one of the first multi-genre sentence-aligned parallel corpus, which also served to benchmark original aligning techniques (Simard, 1998). Today, chunk-aligned Hansards of the 36th Canadian Parliament are freely available for download.

Another key source for reliably translated documents is from the European Union. Its European Commission oversees the *Joint Research Center*, which offers their sentence-aligned parallel corpus JRC-Acquis (Steinberger *et al*., 2012) covering 22 languages (all official European languages except Croatian and Irish) and 231 language pairs totaling 1.5 million documents (1.37 billion words, of which 103,458,996 words are in English) related to debates, press releases, reports, and other parliamentary documents. The European Commission's Directorate-General for Translation also publishes its 24-language 'Translation Memory' (approximately 1.9 million translation units per language) of the legal documents of the Acquis Communautaire (DGT-TM). The original language is not recorded, but 72% were initially drafted in English, before translation by "highly qualified human translators specialized in specific subject domains" and a multi-step verification (Steinberger *et al*., 2012, p. 455), thereby assuring the high quality of both languages, a key factor for language learning.

2.2 Disciplinary contributions

Other major sources of parallel corpora stem from the needs related to human and machine-based translation. Translation memories are databases recording previous translated segments, with the intention of facilitating translations by humans or machines. They may

include essential terminological, linguistic, and contrastive stylistic information. Despite the great wealth of translation memories, Simard (2020) notes that they are "anchored in time and space". Thus, for example, while the French term *pêcheur* was previously translated as *fisherman*, it now calls for the gender-neutral *fisher* (p. 87). While these sources tend to be reserved for private use, exceptions are the DGT-TM (*supra*), MyMemory and the Translation Automation User Society (TAUS), a language industry association that offers a repository of documents provided by members and totalling 35 billion words. However, many of the TAUS documents pertain to machine or automatic translation and may not be pertinent for general English or other domains. Other sources related to language engineering or information technology, which may nevertheless meet certain pedagogical needs, include the European Language Resources Association, the Common Language Resources and Technology Infrastructure (CLARIN) and the mostly fee-based Linguistic Data Consortium (LDC) at the University of Pennsylvania.

Parallel-corpus research also contributes to translation teaching, lexicology, pragmatics, and contrastive analysis. Frérot (2016) describes their use by teachers of translation to target difficulties, as well as for in-class activities, reflecting both academic and professional environments, notably as a complement to translation memory systems. Aijmer (2020) describes the use of parallel corpora for contrastive pragmatics across languages and genres, which attends to a form's function in varying linguistic, social, cultural and historical contexts. These include cognates, such as her study of *absolutely* and the Swedish *absolut*, as well as pragmatic markers, speech acts, information structure and politeness. Drawing upon the field of contrastive analysis and the Europarl corpus, Granger (in press) highlights the interest of n-gram analysis of learner and expert corpora with identifiable translation directions. Identifying differences, she highlights the under- or overuse of lexical bundles in translated or "third code" texts, thereby contributing to our knowledge of both translation

studies and the interlanguage of foreign-language learners, as a 'third code' "arises out of the bilateral consideration of the matrix and target codes: it is, in a sense, a sub-code of each of the codes involved" (Frawley 1984, p. 168).

2.3 Other sources for current research

Literary works, legal documents and movie or television subtitles constitute other reliable sources for parallel corpora. One easy-to-use corpus combining these genres is the Translation Equivalents Database (TREQ), built automatically from the InterCorp parallel corpus including texts in Czech and 27 other languages. The current TREQ interface allows users to select the genre and any of the original languages before translation into either English or Czech, with other languages projected (Škrabal & Vavřín, 2017). Queries may target sequences or include fragments of words, while results are listed by frequency with access to context. The InterCorp corpus (Čermák & Rosen, 2012), containing fictional works, political commentaries, European Parliament documents and subtitles produces "translation candidates" based on unreviewed automatic excerption and word-level alignment results. InterCorp can also be queried via the interface Kontext equipped with various tools according to availability of the original corpus: tokenizers, morphological analysers, taggers and lemmatizers (Čermák, 2019), laying the groundwork for future projects.

Illustrative of the societal need for translated legal documents, Fan and Xu (2002) compiled a 100-file corpus of primarily legal and documentary texts in English (300,000 words) and Chinese (500,000 characters). At the beginning of each sentence, manually inserted hyperlinks permit navigation to a corresponding sentence in the other language. To analyse its usefulness, 63 translation students in Hong Kong were asked to consult the corpus for legal responses related to inheritance and divorce. The legal vocabulary was found to be a veritable hurdle, such as the Chinese literal equivalent to *non-land real property*, deemed more obscure than the English *personal chattels*. The authors conclude, "the corpus provides

instant access to both language versions and since students rely on both languages for comprehension (albeit with Chinese dominant), […] a bilingual corpus of legal language is pedagogically useful" (Fan & Xu, 2002, p. 62). Another larger Chinese-English parallel corpus offering a range of categories (including law, spoken discourse, academic theses) is the downloadable UM-Corpus (15 million aligned sentences) available on the OPUS platform (Tiedemann, 2016) and the Natural Language Processing and Portuguese-Chinese Machine Translation Laboratory (Tian *et al*., 2014).

Another well-represented genre is patent documentation because of international needs. Utiyama and Isahara (2008) compiled a Japanese-English corpus of some 2-million-sentence pairs from patent applications They confirmed that lexical translations were more dependable and easier to treat than long sentences. The Statistical Natural Language Processing Group at the University of Heidelberg created the downloadable Japanese-English BoostCLIR corpus of patent abstracts and German-English or French-English Patent Translation Resource corpora. These patent corpora reflect how financed corporate needs influence the availability of data for international research and student use.

Offering an oasis within the dearth of oral corpora, the public OpenSubtitles repository of movie and television subtitles ranges across many languages and genres, representing slang, narrative and expository discourses. Lison and Tiedemann (2016) treated, notably by algorithms, some 3.36 million-subtitle files representing 2.6 billion sentences distributed across 60 languages before making them available via the OPUS platform. Subtitles, in blocks of 50 characters maximum, were aligned using timing information, plus other corrections and metadata to assure quality.

Finally, academic parallel corpora are rare, but are of considerable interest to teaching language in higher education. A recent addition is the English-French, syntactically annotated ParaSHS corpus incorporating some 1,000 research articles in the Humanities (Kraif, 2018).

3. Tools and corpora defining core issues for teaching and learning

More so than for other corpus studies, core issues are defined in function with query tools. The underlying or accessible tools largely influence the type and results of a query. Parallel corpora resources announce their document sources and tools, which is not necessarily the case for well-known resources such as Linguee, ReversoTranslation and WordReference. Rabadán (2019, p. 59) notes that they often lack information on the language data or the tools serving for alignment or reference to frequency. These absences may affect the reliability of query results, especially as related to frequency and genre (Hartwell, 2020), which is not the case for parallel corpora. To illustrate how tools and document availability condition teaching and learning possibilities, we now examine some key sources.

3.1 Tools

Queries rely upon tools common to analyzing many of the same elements as monolingual corpora: keyword in context, frequency lists, collocates, search grammars. There are several levels of query: immediate consultation of an online interface, downloading a corpus for query with a separate tool, or creating one's own corpus using free aligning tools, such as Hunalign, LF aligner or Web Align Toolkit. These tools offer a predictive power for cross-language consultation surpassing the word-to-word translations often found in dictionaries.

One of the most advanced on-line user-friendly sources is OPUS, a database of 34 languages in a range of genres, as well as a collection of downloadable bilingual or multilingual corpora. It relies upon the open-source architecture Corpus Workbench, designed at the University of Stuttgart and used for the British National Corpus, later influencing Sketch Engine (Evert & Hardie, 2011). On OPUS, users first select either the aligned-sentence interface or the lexical correspondence interface, built upon translation

probability and parallel concordancing. After selecting a corpus, such as OpenSubtitles2018, Tatoeba or TedTalks, then the source and target language, users may query an individual item or sequence, lemma or part of speech. Thus, users can relatively easily access complex language data within a specific genre or field.

In contrast, other sources must be downloaded. The European Parliament includes a downloadable tool to query their parallel corpus. Another option is Sketch Engine, requiring a paid individual or institutional subscription, allows users to query a parallel corpus acquired elsewhere, such as those found on OPUS, via the "Create Corpus" option. Other, more targeted initiatives continue to be created, such as the 280,000-word sentence-aligned European Parliament Translation and Interpretation Corpus (EPTIC) integrating cleaned transcriptions, interpretations, translations and additional uncleaned transcriptions (including metadata) of parliamentary speeches in English, French and Italian, as well as access to time-synced video documents (Ferraresi & Bernardini, 2019). The EPTIC corpus illustrates the possibilities offered by advanced uses of tools and available documents.

3.2 Language availability

Another key factor determining core issues is the availability of quality language translations. Although both OPUS and CLARIN (86 downloadable multilingual parallel corpora) declare the objective of incorporating a wider range of languages, documents in European languages continue to constitute the bulk of available data. Other projects seek to compile bilingual corpora with often diverging alphabets or scriptural traditions, which raises new questions of automatic analysis, tagging or alignment. For instance, languages such as Arabic or Chinese need dedicated tools for word and sentence segmentation, as blank spaces, capital characters and punctuation marks do not constitute reliable clues for these tasks.

Mikhailov and Cooper (2016) detail some thirty parallel corpora, most of them including English. Among these are the publicly available Amsterdam Slavic Parallel Aligned Corpus of literary texts, the Svrokorpus collection of Slovene-based bilingual texts, the Multilingual Corpora for Cooperation (9 European languages) of comparable financial newspapers plus the *Journal of the European Commission*, the paragraph-aligned Multilingual Corpus of Legal Documents Corpus (English, Finnish, Swedish, Russian) and the Russian National Corpus including literary classics and their translations.

To remedy the lack of representation of certain languages, one project underway is the King Saud University's 10-million word Arabic–English Bidirectional Parallel Corpus, targeting eight themes, from biographical, medical or scientific. Manually compiled, cleaned, and aligned at the sentence level, one of the major hurdles has been the lack of programs capable of compiling such resources (Alotaibi, 2017).

Within a context of increasingly present language technology, many African languages remain digitally underrepresented. However, the part-of-speech tagged, lemmatized and sentence-aligned 1.2-million-word downloadable Swahili-English SAWA Corpus draws upon dictionary exemplars, the Kenyan Constitution, Kenyan investment reports, movie subtitles, non-governmental organization leaflets, United Nations documents and religious texts (De Pauw *et al*., 2011). Difficulties resulting from the morphologically more complex nature of Swahili were lessened by morphologically deconstructing Swahili words (De Pauw *et al*., 2011, p. 337).

Another response to enlarging language availability in addition to English is Tatoeba, a collaborative and open collection of sentences and their translations into a multitude of languages, each sentence tagged with the contributor' name. Some 365 languages are supported, including over 37,000 sentences in Toki Pona, 22,000 in Persian and 3,000 in Thai. Attention is given to source, for example the word *Enkosi!* ("Thank you" in Xhosa) is

listed as initially translated into four languages, before translating these for some 141 new languages. Although without a precise research objective, the site is remarkably clear and easy-to-use, as well as being controlled for quality, all of which are key issues for language learning.

3.3 Current contributions to teaching and learning

The access to tools and languages are initial considerations for any teacher wishing to query parallel corpora. As for unilingual corpora, classroom applications may favor data-driven learning (DDL). Illustrating a "good example" of DDL, Cobb (2019, p. 195) cites Chan and Liou (2005) who asked students to complete gap-fill exercises by querying a Chinese-English parallel corpus in order to generalize about language patterns, typically verb-noun collocations, such as *set fire*. Further illustrating computer-assisted language learning, Johns *et al.* (2008) found that incorporating Chinese translations into the corpus consulted by a Taiwanese secondary school English literature class allowed students to rely on their first language for comprehension, thereby surpassing reductive word-to-word translation.

Both direct computer-based and indirect paper-based queries contribute to DDL. Chujo and Oghigian (2012) drew upon a Japanese-English newspaper corpus, comprised of 150,000 aligned translation pairs accessible via the concordancer Paraconc, to propose both computer concordancing and paper concordance data to Japanese engineering students learning basic English. Students discovering both mediums showed greater capacity to identify and to produce noun and verb phrases, than the control group. Overall, approximately three-quarters of the students declared that the computer-based DDL 1) offered more necessary translations, 2) provided a sensation of security, 3) helped to translate a specific sentence, and 4) helped to grasp word meaning compared to half or less of the students who found this true for the paper-based DDL translations. The authors suggest

adopting a blended approach beginning with paper-based exercises to control for focus before introducing computer-based DDL to reinforce learning. They conclude that "students faced with non-vetted computer-based DDL use the parallel translation to confirm the meaning so that they can focus on the grammatical structure" (Chujo & Oghigian, 2012, p. 180). Thus, learning can be enhanced by alternating between paper and numeric supports and also by alternating attention between form and meaning.

Friginal (2018) details a teaching unit in which pre-Intermediate Japanese students of English consult the sentence-aligned Japanese-English WebParaNews Corpus of news articles from a bilingual newspaper to notice, through hands-on concordancer queries, the different uses of synonyms such as *start/begin* and *big/large*. Then, students edited their own written production by replacing certain words with more "natural-sounding" ones using the same concordancer techniques. Once students are comfortable employing these techniques on the user-friendly, but smaller WebParaNews Corpus, they move on to a more complex, tool-equipped unilingual corpus.

Frankenberg-Garcia (2012a) explains "it takes time and substantial training to become a proficient corpus user, but learners needn't become experts in corpus linguistics. Simple demonstrations of how corpora can be utilized to answer authentic questions that emerge in class will do" (p. 50). In 15 minutes, she created a paper handout of Compara and OpenSubtitles concordance data for students to examine the Portuguese *segurança,* equivalent to either "security" or "safety", allowing them to identify contextual differences mandating use. Verification of lexical use is one of the most accessible benefits of parallel corpora.

The difficulty of identifying lexical differences between languages is also highlighted by the authors of the sentence-aligned Russian Learner Translator Corpus (2.3 million English or Russian tokens) produced by translation students from 14 Russian universities.

This corpus can be queried for individual items or sequences, lemmas or parts of speech and covers ten genres, including news reports, letters and interviews. They suggest investigating challenging English lemmas lacking immediate counterparts in Russian such as *overqualified* or *lock-in*, and *faux amis*, such as *actual*, *decade* or *economical*. In another example of confusing multiple lexical translations, Frankenberg-Garcia and Santos (2003) queried the English-Portuguese parallel Compara Corpus to help students understand that *actualmente* does not translate to the English cognate "actually", but rather terms such as "now", "nowadays", "at the moment" (pp. 387-388). These current contributions serve as foundations for future projects and contribute to better teaching practices.

4. Case Studies

Building upon these studies, we now explore the pertinence of parallel corpora in advanced language learning environments through three case studies. The first study illustrates the importance of the resource for identifying accurate meaning, the second examines false cognates across multiple parallel corpora, and the third delves into the capacity to examine complex constructions.

4.1. Case study 1: Existing multilingual resources

This analysis illustrates how the data of a resource, whether general data from the Internet or from specific corpora, influences the quality of lexical resources. We compare several well-known on-line dictionaries, Linguee, ReversoTranslation and WordReference, which Doval and Sánchez Nieto (2019, p. 3) define as dictionaries "enlarged with multilingual online resources" and one parallel corpus platform, Tradooit.

In order to test the reliability of these sources, we analysed proposed translations of the expression *third degree* into French. The two terms 'third' and 'degree' are also

frequently found together to describe a burn, an assault or a status in computer science. However, the expression *third degree* takes on a specific meaning in a legal context, as found in the United States Supreme Court's opinion for Miranda v. Arizona (1966), "it is clear that police violence and the 'third degree' flourished [during the 1930s]". According to the *Merriam-Webster dictionary*, 'third degree' refers only to "the subjection of a prisoner to mental or physical torture to extract a confession". The *Cambridge dictionary* which also has an entry for 'third-degree burn', considers it to be "asking serious questions and/or giving someone rough treatment to get information".

Linguee proposes *troisième degré*, but also *troisième diplôme*, building upon 'degree' as a university diploma. Wordreference also first refers to medical burns, followed by 'getting and giving the third degree' with the weaker translation *interroger* ("to interrogate") and the idiomatically appropriate *cuisiner* ("to cook"). ReversoDictionary offers the medically-based example "third degree atrioventricular block". ReversoContext highlights two possible uses, *troisième degré* and *interrogatoire* ("interrogation"). *Troisième degré* is illustrated by a reference to the severity of burns and the example "No wonder I'm getting the third degree", which does not aid understanding. In ReversoContext, *interrogatoire* ("interrogation") is contextualized by the bilingual example (1), in which the 'suggested translation' (ST) in French, containing the ambiguous *interrogation*, is clarified by the adjective *severe*, as reflected in the 'literal translation' (LT), included here for greater comprehension of the French counterparts.

(1) The suspect was given the third degree until he confessed his crime.

(ST : "Le suspect avait eu un interrogatoire sévère avant qu'il n'avoue son crime" (ReversoContext))

(LT: "The suspect had a severe interrogation before confessing his crime")

Tradooit, one of the rare sources to offer the option to select and identify corpus data, proposes the translations *troisième degré*, *interrogatoire* and *inquisiteur* ("inquisition") as well as the relevant *passage a tabac* ("a violent beating") and *cuisinage* ("to cook"). In contrast, the vast majority of TAUS references apply to informatics or computer science, such as "Does not export entities as third-degree B-splines". Finally, the OpenSubtitles Corpus offers 151 English-French matches, many of which apply to policing: *interrogatoire musclé* ("muscled interrogation") and *les flics m'ont donné toutes sortes de traitement de choc* ("the police made me undergo all sorts of shock treatments").

These resources are easy to use and accessible on a mobile device or computer. However, as we have seen, the results are only as reliable as the quality of the language of the corpora they draw upon. In other words, a source that relies heavily upon general English may not offer the formal language required for academics. The capacity to deal with multi-word sequences or the verification by community members are both important. The subject domain of the corpus is determinant, as, for example, noted for the computer-science oriented TAUS. In the Supreme Court opinion, "third degree" was within quote marks to signal its informal character. This informality is characteristic of fictional films and television programs, as exemplified in the results found here. The results for *third degree* exemplifies how the qualities of the data influence results. General language data does not always respond well to specific needs and may actually mislead scholarly users.

4.2 Case study 2: False cognates

Drawing upon Frankenberg-Garcia and Santos (2003), we queried three corpora for several well-known French/English false cognates: the online interfaces Tradooit and TREQ, plus the downloaded GlobalVoice available on OPUS and queried via Sketch Engine. By comparing the translation results from French to English of the "false friends", *actuellement*, *demande*,

évidence and *réalisation*, we hope to illustrate the technical and corpus differences of these options. TradooIT allow users to choose from a range of corpora (Europarl, UN, subtitles, etc.). Here, we have selected the entire corpora, which offers higher frequency, but without attention to genre.

All three sources give simultaneous "keyword in context" (KWIC) results, simultaneously listing both the French and English equivalents. TradooIT only searches one word form at a time, here the keyword's singular form. TREQ and Sketch Engine are case sensitive and thus show frequencies for *obviously* and *Obviously*. They all offer the option to view frequency. In order to access frequencies via Sketch Engine, the user must first search for the keyword in the source language, identify the recurrent translations in the target language and include these in a second query. Frequencies can be viewed according to word form, lemma or part of speech. Table 31.1 lists the top seven translations and their frequency by word form, although actual query results may include supplementary data.

The results vary according to the corpora database and the tools, as can be understood by observing the results. For *actuellement,* GlobalVoices has a greater frequency of the culturally-associated equivalent *moment*, also found in the Cineurope and the OpenSubtitles Corpus integrated into TradooIT. Although *request* or *ask* are frequent equivalents of *demande, application* is listed by both TradooIT and TREQ, due to their incorporation of governmental sources. TradooIT queries respect singularity or plurality, but not capitalization, thereby giving rise to less repetition (within the first seven results) of target-words often beginning a sentence, such as *évidence.* In contrast, GlobalVoices and TREQ list both *obviously* and *Obviously*, splitting the frequencies of this term. Finally, for *realization, conduct* occurs notably within TradooIT's incorporation of documents from the Canadian Nuclear Safety Commission, while TREQ's *attainment* originates primarily from the Acquis

Communautaire corpus. Thus, these frequency results may help learners explore a range of unexpected equivalents, especially according to genre.

Table 31.1 Lexical comparison

	GlobalVoices	***TradooIT***	***TREQ***
actuellement	now (6421), current (1725), Now (1322), currently (1090), moment (907), present (848)	currently (21504), now (6922), is (5384), current (2819), present (1347), presently (1250), today (1027)	currently (4016), now (723), present (599), current (407), moment (237), today (118), being (117)
demande	asks (255), demand (209), wonder (154), request (15830), wonders (141), asked (131), demands (119)	demand (16819), request (15830), asked (10544), application (9871), wonder (3783), calls (2910), apply (2570)	request (9056), application (7331), demand (3466), ask (1543), applications (1023), wonder (922), calls (805)
évidence	obviously (44), clearly (33), Obviously (27), obvious (14), highlights (14), highlight (13), evidence (11)	obviously (7501), clearly (4024), highlight (2697), evidence (1089), obvious (817), shows (817), identified (388)	obviously (220), highlighted (152), clearly (134), highlight (97), Obviously (95), Clearly (88), obvious (88),
réalisation	achievements (38), project (10), achievement (8), implementation (7), carry (3), realisation (3), completion (2), achieved (2), achieve (2)	achievement (2825), achieving (2676), project (2415), carrying (884), delivery (829), conduct (815), implementation (731)	achievement (731), achieving (463), completion (337), implementation (323), realisation (212), achieve (193), attainment (182)

4.3 Case study 3: Complex phenomena

Parallel corpora also offer possibilities to examine complex and non-intuitive phenomena, which are of direct interest to language learners. For example, Frankenberg-Garcia (2012b) relates a study concerning the appropriateness of ending a business letter with either *I look forward to hearing from you* or *I am looking forward to hearing from you.* A query of the one-million-word English-only Business Letter Corpus of American and British letters confirmed that both are acceptable, but that the former is more conventional. It is possible to

undertake such an investigation thanks to a parallel corpus. To illustrate the technical possibilities, we introduce here the results of a sequence query.

Learners often have difficulty grasping when past tense forms should be used. This is true for francophone learners because a present tense form is used in French even when evoking events beginning in the past. Furthermore, in English, time can be introduced by either *for* or *since*. The interest of consulting a parallel corpus for differentiating between *since* or *for* is also highlighted on the Russian Learner Translator Corpus platform. In French, both of these prepositions are translated as "depuis", thus rendering a direct word-to-word translation impossible. To help learners notice differences in French, a query of the sequence containing a first-person pronoun, a verbal part of speech (pos), and the preposition *depuis* ("since/for") could take the form of Example 2:

(2) je [pos=V.*] depuis

LT : ("I [pos=V.*] since|for")

A query of Europarl.v7 on OPUS produces 86 matches, with 36 varying lexical verbs. These sequences show repeatedly that while the French verbs are in the present tense, the English versions take a past tense form. In Example 3, the lexical verb *faire* takes a singular present tense form: *fais* ("do/make"). However, the proposed equivalent is *have been doing*, as a simple present tense would be incorrect in English.

(3) Peut-être le mieux est-il encore de s'en tenir à ce que **je fais** *depuis* quinze ans

(LT: "Maybe the best is still to carry on as **I do** *for* 15 years")

(ST: "Perhaps the best all-round solution is to carry on as **I have been doing** *for* the past 15 years" (Europarl.v7))

Example 4, also from Europarl.v7, displays multiple non-intuitive phenomena of a French sequence. First, the more frequent French ambiguous pronoun of the collocation *on sait* ("one knows") serves to generalize the source of the information by its inclusive nature

(Hartwell & Jacques, 2014) and is translated here by “as you all know”. Once again, the central lexical verbal collocation *je plaide* (“I plead”) is in the present tense in French, but takes the present perfect continuous tense in English, “have been calling for”.

(4) **On** sait que **je plaide** depuis…

(LT: “**One** knows that **I plead** for/since…”)

(ST: “As you all know, **I have been calling for**…” (Europarl.v7))

Example 5 confirms that the equivalent of a French present tense *souligne* (“underline”) is a present perfect continuous tense in English, more often translated as “have been highlighting”. Also notable for English language learners is the translation of *depuis* to “since” instead of “for”, followed by a date. Learners can benefit from seeing that *since* is followed by a date, but *for* by a period of time.

(5) Il subsiste certaines préoccupations, **que je souligne depuis** 1999.

(LT: “It exists certain concerns that I **underline since** 1999.”)

(ST: “There are still areas of concern that **I have been highlighting since** 1999.” (Europarl.v7))

A query of the GlobalVoices Corpus on OPUS of Example 2, results in 20 occurrences. Among the pertinent ones, example 6 offers a more complex translation of the French *je fais depuis* (LT: “I do/make for”), by adopting a new grammatical category: “the critique that I make” becomes simply “my critique”. Furthermore, *depuis longtemps* (“for a long time”) becomes the compact *long standing.*

(6) qui confirme la critique que **je fais depuis** longtemps des arts plastiques en Jamaïque

(LT: “which confirms the criticism that **I make for** a long time about plastic arts in Jamaica”)

(ST: “which bears out **my long-standing** critique of visual art in Jamaica”

(GlobalVoices)

Thus, these examples allow language learners to visualize the multiple ways of formulating meaning in different languages. Language rules, such as the use of a present or past tense, become more explicit to language learners as they view a compiled set of occurrences. Corpus queries offer condensed collections of input, thus illustrating rules and making patterns more apparent. Reliable bilingual data allow learners to concentrate on language differences or patterns, with less cognitive load related to comprehension.

5. Recommendations for teaching and learning

5.1 Caveats of corpora

Many of the caveats for teaching and learning with parallel corpora are similar to those decried for corpora in general, notably, the necessary time, training, technical and ergonomic accessibility. As Frankenberg-Garcia (2016) notes, mastering how to select a corpus or to query and interpret concordances, word lists, collocations or other data is a first step. Transposing this expertise to the classroom is a second, for “corpus-based teaching aids must be relevant, useful and accessible to the particular group of learners” (Frankenberg-Garcia, 2016, p. 394).

However, these observed difficulties are exasperated by reduced availability of parallel corpora and, for some queries or corpora, increased technical needs. Students tend to access certain common resources, such as Linguee, despite the existence of free corpus-based and user-friendly sites, such as TradooIT. The current shortages of existing parallel corpora depend greatly on research funding benefitting or not a given language community. However, community endeavours, such as GlobalVoices and Tatoeba, are new conduits between less represented languages and English language learning.

5.2 Simplification and over-normalization

Another weakness of calling upon parallel corpora is the issue of the quality of the target language found in translated texts and the phenomenon known as “translationese” (Aijmer, 2020; Baker, 1998). Quality translations stretch beyond word-to-word translations to adopt more complex equivalents, such as a change in grammatical category (Example 6). This is why many parallel corpora are aligned at the sentence level instead of by word. Understanding subtle connotation is also a learning challenge, as Kübler (2011) highlights, for example as related to the neutral *to cause* and the French *causer*, the latter introducing a negative result. She suggests that consulting corpora and specifically specialized corpora may help students to “avoid using *causer* as the translation equivalent of those English verbs of causation that do not have a negative semantic prosody” (Kübler, 2011, p. 77).

Furthermore, relying upon unique data sources of corpora increases the risk of an unwanted normalization, such as “eurolect”, a manifestation of converging terminology and linguistic interferences within European Parliament documents (Torrellas Castillo, 2009). Cultural, social and political histories influence subject matter and the associated discursive patterns across communities. For English language learning, the student population as well as their academic and professional discursive needs should also be taken into account. Kubota and Chiang (2013) confirm that “it is necessary to explore contextual understanding of *needs* by taking into consideration how learner’s gender, race, class, and other backgrounds shape social practices in a specific professional context” (p. 495). Thus, teaching activities and materials incorporating parallel corpora should build upon language learners’ diverse needs as related to learners’ first language, with attention to the corpora’s underlying content and socio-political positioning. Corpora, like language, is not neutral, meriting a teacher’s attention to possibly problematic content or the normalization of diverse learners’ needs.

6. Future directions of research

Machine translation tools provide solutions based on corpora and statistical probability models. Systems incorporating machine translation tools offer an indirect way of accessing parallel corpora, as these systems essentially rely on existing translation corpora. Until the early 2010's, Statistical Machine Translation systems (Koehn *et al*., 2003) suggested a translation by coupling word- or phrase-level translation probabilities and probabilities linked to the model of the target language. This coupling improves the idiomatic character of the suggested translations. The knowledge introduced by these language and translation models were, to some extent, made explicit by the corresponding probability measures.

In more recent models based on deep-learning techniques (Bahdanau *et al.*, 2015), sentences are translated, not by combining translation fragments, but on the basis of a global representation of their meaning, thereby offering significant improvements especially as related to the idiomaticity. However, accessing the information encoded by the neural network remains difficult, as the parallel corpora feeding the abstract network is not always easily identified. As we have seen, the source data is essential to the quality of the results.

In the future of artificial intelligence, modulating the suggested translations according to the textual genre of the parallel texts would be an important step forward. For example, Chambers (2010) comments that the 87 occurrences of the verb *connaître* (commonly, "to know") in the Chambers-Le Baron Corpus of French research articles refer to "doing an experience". Thus, a unilingual corpus helps to identify meaning often within a specific genre, but does not suggest translations, as does a parallel corpus such as the ParaSHS English-French parallel corpus of research articles in the humanities (Example 7). Here, *connaîtrons* is translated as "experiencing", which mirrors Chamber's (*ibid*) understanding of the word in an academic context.

(7) **Connaîtrions**-nous alors une mutation des liens entre **moi** et corps et, corrélativement, une mutation de l'imaginaire tel que le définit Lacan ? (ParaSHS)

(ST: Are we thus **experiencing** a change in the relations between the **self** and the body and, by correlation, of the imaginary as Lacan defines it? (ParaSHS)

The Tradooit automatic neuronal translator based on Canadian governmental documents also adopts of form of ("experience") to translate the verb *connaître* (Example 7). However, it proposes ("me") instead of the psychological notion of "self".

(8) Would we then **experience** a mutation in the bonds between **me** and body and, consequently, a mutation of the imagination as defined by Lacan? (Tradooit)

If Deepl (Example 8) correctly identifies *moi* as the ("self"), it suggests the frequent general English verb ("know"), which does not ring true.

(9) (ST: Would we then **know** a mutation of the links between **self** and body and, correlatively, a mutation of the imaginary as defined by Lacan? (Deepl))

Thus, parallel corpora and automatic translation systems evolve synchronically. Automatic translations rely upon the quality and type of data. Parallel corpora expand according to the evolution of tools as well as the academic, societal, corporate capacities and projects.

Technological advances corresponding to greater accessibility and ease-of-use – as well as teacher training in corpus use – will influence the future of parallel corpora for language teaching. Actual academic and professional practices of consulting these resources should modify their introduction and role in the classroom as a strategy for lifelong learning. The acquisition of core vocabulary or grammatical understanding should accompany the capacity to consult critically and successfully available resources, such as parallel corpora. This can be done by combining direct consultation or indirect study from chosen output. For non-native learners of English, the capacity to consult a set of corresponding English, their

language occurrences targeting a specific lexical or grammatical question is a great advantage especially when adapted to language level. This advantage complements and enforces theoretical explanations or decontextualized data. Consulting parallel corpora can focus on meaning and form. Thus, understanding teacher, learner, material designers and professional practices of consulting parallel corpora is another area for future research.

Further reading

Doval, I., & Sánchez Nieto, M.T. (2019). *Parallel corpora for contrastive and translation studies*. Benjamins.

This collective work offers an overview of parallel corpora, notably for translation.

Fan, M., & Xu, X. (2002). An evaluation of an online bilingual corpus for the self-learning of legal English. *System*, *30*, 47–63.

A study confirming the pedagogical interest for translation students to consult a Chinese-English legal corpus containing navigational hyperlinks to corresponding sentences in the other language.

Frankenberg-Garcia, A., & Santos, D., (2003). Introducing COMPARA the Portuguese-English parallel corpus. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 71–87). St. Jerome.

One of several articles about the groundbreaking Portuguese-English parallel corpus and of the many uses of parallel corpora for teaching and learning applicable across languages.

Corpus and interface links

Bitextes anglais-français corpus: http://rali.iro.umontreal.ca/rali/?q=fr/BAF

Canadian Parliament Hansards: https://www.isi.edu/natural-language/download/hansard/

Common Language Resources and Technology Infrastructure: www.clarin.eu/content/language-resource-inventory

Compara: www.linguateca.pt/COMPARA

Deepl: https://www.deepl.com

EIIDA: https://corpora.aiakide.net/scientext20/?do=SQ.setView&view=corpora

English-Norwegian Parallel corpus: https://tekstlab.uio.no/glossa2/saml?licence=ACA-NC-LOC-LRT-ND_OMC;back=https%3a%2f%2ftekstlab.uio.no%2fglossa2%2fomc4

European Language Resources Association: www.elra.info/en/about/elra/

European Parliament corpus: https://ec.europa.eu/jrc/en/language-technologies/dcep

European Commission's Translation Memory: https://data.europa.eu/euodp/en/data/dataset/dgt-translation-memory

European Research Infrastructure Consortium: https://www.clarin.eu/resource-families/parallel-corpora

Linguistic Data Consortium: www.ldc.upenn.edu/about

MyMemory: https://mymemory.translated.net/

OPUS: http://opus.nlpl.eu/

ParaSHS: http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0

Russian Learner Translator Corpus: https://rus-ltc.org/static/html/about.html

Sketch Engine: https://www.sketchengine.eu/guide/setting-up-parallel-corpora/

Statistical Natural Language Processing Group: https://www.cl.uni-heidelberg.de/statnlpgroup/

United Nations corpus: https://conferences.unite.un.org/UNCORPUS/en/DownloadOverview

Tatoeba: https://tatoeba.org/eng

Translation Automation User Society: https://data-app.taus.net/

Translation Equivalents Database: http://portal.clarin.nl/node/18403

UM-corpus: http://nlp2ct.cis.umac.mo/um-corpus/index.html

Web Align Toolkit: http://phraseotext.univ-grenoble-alpes.fr/webAlignToolkit

References

Aijmer, K. (2020). Contrastive pragmatics and corpora. *Contrastive Pragmatics*, *1*, 28–57. https://doi.org/10.1163/26660393-12340004

Alotaibi, H.M. (2017). Arabic-English parallel corpus: a new resource for translation training and language teaching. *Arab World English Journal*, *8*(3), 319–337. https://dx.doi.org/10.24093/awej/vol8no3.21

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). *Neural machine translation by jointly learning to align and translate* [Paper presentation]. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, United States. https://arxiv.org/pdf/1409.0473.pdf

Baker, M. (1998). Réexplorer la langue de la traduction: une approche par corpus. *Meta: Translators' Journal*, *43*(4), 480–485.

Carter-Thomas, S., & Jacques, M.-P. (Eds.). (2017). *CHIMERA. Romance Corpora and Linguistic Studies*, *4*(1). https://revistas.uam.es/index.php/%20chimera/article/view/6948

Čermák, F. (2019). InterCorp: a parallel corpus of 40 languages. In I. Doval & M.T. Sánchez Nieto (Eds.), *Parallel corpora for contrastive and translation studies* (pp. 93–102). Benjamins.

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427.

Chambers, A. (2010). L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé. *Revue française de linguistique appliquée (Vol. XV)*, 2010/2, 9–20.

Chan, T. P., & Liou, H. C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations. *Computer Assisted Language Learning, 18*(3), 231–251.

Chartier, R. & Martin, H.-J. (1986). *Histoire de l'édition française, 4*. Fayard.

Chujo, K., & Oghigian, K. (2012). DDL for EFL beginners: a report on student gains and views on paper-based concordancing and the role of L1. In J. Thomas & A. Boulton (Eds.), *Input, process and product: developments in teaching and language corpora* (pp. 169–182). Masaryk University Press.

Cobb, T. (2019). From corpus to CALL: the use of technology in teaching and learning formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 192–210). Routledge.

De Pauw, G., Wagacha, P. W., & de Schryver, G.-M. (2011). Exploring the SAWA corpus: Collection and deployment of a parallel corpus English-Swahili. *Language Resources & Evaluation, 45, 331*–344. https://doi.org10.1007/s10579-011-9159-7

Ernst-Slavit, G., & Mulhern, M. (2003). Bilingual books: promoting literacy and biliteracy in the second language and mainstream classroom. *Reading Online, 7*(2), 1–15.

Evert, S., & Hardie, A. (2011). *Twenty-first century corpus workbench: updating a query architecture for the new millennium* [Paper presentation]. Corpus Linguistics 2011, University of Birmingham, UK. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf

Ferraresi, A., & Bernardini, S. (2019). Building EPTIC: A multi-sided, multi-purpose corpus of EU Parliament proceedings. In I. Doval & M. T. Sánchez Nieto (Eds.), *Parallel corpora for contrastive and translation studies* (pp. 123–139). Benjamins.

Frankenberg-Garcia, A. (2012a). Integrating corpora with everyday language teaching. In J. Thomas & A. Boulton (Eds.), *Input, process and product: developments in teaching and language corpora* (pp. 36–53). Masaryk University Press.

Frankenberg-Garcia, A. (2012b). Raising teacher's awareness of corpora. *Language Teaching, 45*(4), 475–489.

Frankenberg-Garcia, A. (2016). Corpora in the classroom. In G. Hall (Ed.), *Routledge handbook of English language teaching* (pp. 383–398). Routledge.

Frérot, C. (2016). Corpora and corpus technology for translation purposes in professional and academic environments. Major achievements and new perspectives. *Cadernos de Tradução, 36*(1), 36–61. https://doi.org/10.5007/2175-7968.2016v36nesp1p36

Friginal, E. (2018). *Corpus linguistics for English teachers: New tools, online resources, and classroom activities.* Routledge.

Granger, S. (in press). Tracking the third code: A cross-linguistic corpus-driven approach to metadiscursive markers. In A. Čermáková & M. Mahlberg (Eds.), *Corpus as Discourse*. Benjamins.

Hartwell, L., & Jacques, M.-P. (2014). Authorial presence in French and English: Pronoun + verb patterns in Biology and Medicine research articles. *Discourse, 15.* https://doi.org/10.4000/discours.8941

Hartwell, L. (2020). A didactic comparison of online French-English lexical resources. *Études en Didactique des Langues*, *34*, 7–24.

Johansson, S. (2008). *Contrastive analysis and learner language: A corpus-based approach*. University of Oslo. https://www.hf.uio.no/ilos/forskning/grupper/English_Language_and_Corpus_Linguistics_Research/papers/contrastive-analysis-and-learner-language_learner-language-part.pdf

Johns, T., Lee, H.-C., & Wang, L. (2008). Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning*. *21*(5), 483–506.

Koehn, P., Och, F.-O., & Marcu, D. (2003). Statistical phrase-based translation. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL): Proceedings*, Volume 1 (pp. 48–54).

Kraif, O. (2018). Constitution et traitement d'un corpus bilingue d'articles scientifiques : exemple de mise en oeuvre automatique avec une architecture légère en Perl. In M. Mangeot & A. Tutin (Eds.), *Actes des Journées LTT 2018*, Grenoble.

Kübler, N. (2011). Working with different corpora in translation teaching. In A. Frankenberg-Garcia, L., Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 62–80). Continuum.

Kubota, K., & Chiang, L. T. (2013). Gender and race in ESP research. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 481–499). Wiley-Blackwell.

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th international conference on language resources and evaluation*. https://www.semanticscholar.org/paper/OpenSubtitles2016%3A-Extracting-Large-Parallel-from-Lison-Tiedemann/e11edb4201007530c3692814a155b22f78a0d659

Mikhailov, M., & Cooper, R. (2016). *Corpus linguistics for translation and contrastive studies: A guide for research.* Routledge.

Rabadán, R. (2019). Working with parallel corpora: Usefulness and usability. In I. Doval & M. T. Sánchez Nieto (Eds.), *Parallel corpora for contrastive and translation studies* (pp. 57–78). Benjamins.

Škrabal, M., & Vavřín, M. (2017). The Translation Equivalents Database (Treq) as a lexicographer's aid. In I. Kosek, *et al.* (Eds.), *Proceedings of the eLex 2017 conference: Electronic lexicography in the 21st century* (pp. 124–137).

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 454–459), Istanbul.

Simard, M. (2020). Corpora in the classroom. In M. O'Hagan (Ed.), *Routledge handbook of translation and technology* (pp. 78–90). Routledge.

Simard, M. (1998). The BAF: A Corpus of English-French Bitext. *Proceedings of the First International Conference on Language Resources and Evaluation* (pp. 489–494), Granada, Spain. http://www.mt-archive.info/LREC-1998-Simard.pdf

Teubert, W., & Čermáková, A. (2004). Directions in corpus linguistics. In M.A.K. Halliday, (Ed.), *Lexicology and corpus linguistics* (pp. 113–165). A&C Black.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., Li, S., Wang, Y., & Lu, Y. (2014). UM-Corpus: a large English-Chinese parallel corpus for statistical machine translation. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Tiedemann, J. (2016). OPUS - Parallel corpora for everyone. *Baltic Journal of Modern Computing, 4*(2), 923–926.

Torrellas Castillo, M. (2009). *Les interférences linguistiques dans les textes en espagnol des institutions de l'Union Européenne : Étude fondée sur le corpus bilingue massif*

aligné de l'acquis communautaire. [Unpublished doctoral dissertation]. Université de Poitiers.

Utiyama, M., & Isahara, H. (2008). Mining patents for parallel corpora. In C. Goutte, N. Cancedda, M. Dymetman, & G. Foster (Eds.), *Learning Machine Translation.* https://direct.mit.edu/books/chapter-pdf/228866/9780262255097_cab.pdf

Laura M. Hartwell is full professor of English studies at the University Toulouse Capitole and Director of the Lairdil laboratory in Toulouse, France. Her research interests include numerical language resources and English for legal purposes. She obtained a French HDR diploma at the University of Grenoble Alpes (Lidilem) and a Doctoral degree at the University of Toulouse (Lairdil), both on academic English and its teaching.

Olivier Kraif is full professor in Grenoble at the Université Grenoble Alpes (UGA) and a member of LiDiLEM laboratory. He teaches in the fields of Computer Science, Computational linguistics and Natural Language Processing. He has a specific interest in text corpora processing, especially related to multilingual corpora (comparable as well as parallel). His research aims at developing techniques and tools to investigate linguistic phenomena from various points of view: lexicon, phraseology, contrastive analysis and translational studies.