Granger, S. & M.-A. Lefer (2022). Corpus-based translation and interpreting studies: A forward-looking review. In S. Granger & M.-A. Lefer (eds), *Extending the Scope of Corpus-based Translation Studies*. Bloomsbury Advances in Translation series. London: Bloomsbury, 13-41.

Chapter 1

Corpus-based translation and interpreting studies: A forward-looking review

Sylviane Granger and Marie-Aude Lefer

University of Louvain

The aim of the chapter is to offer a forward-looking review of corpus-based translation and interpreting studies. With a view to providing an objective panorama of recent trends in the field, we present a research survey based on 186 corpus studies published in English in twelve top-rated translation and interpreting journals between 2012 and 2019. The corpus studies included in the survey all meet two requirements: they rely on corpora available in electronic format and make use of corpus-linguistic techniques and tools to analyse them. In contrast to previous surveys, which are mostly based on bibliometric records, our survey relies on a detailed exploration of the full texts of the articles. In the chapter, we describe the major trends that emerge in three categories of corpus studies: theory- and methodology-oriented studies, applied studies and empirical studies. In the case of empirical studies, which account for approximately two thirds of our dataset, we analyse in detail a range of specific aspects, such as research focus (linguistic focus and translation features), corpus design (corpus types, corpus size, modality and register, languages), corpus techniques and statistical testing. The survey provides a wealth of insights into the current status of the field, testifying to the growing maturity of corpus research in translation and interpreting studies while also identifying areas where progress has been relatively modest. The chapter ends with some forward-looking suggestions for the field.

Keywords: corpus-based translation studies, corpus-based interpreting studies, survey, corpus methodology, corpus applications, linguistic focus, translation feature, corpus design, corpus techniques, statistical testing

1. Introduction

The origin of corpus-based translation studies (CBTS) can be traced back to an article entitled 'Corpus Linguistics and Translation Studies: Implications and Applications' (1993). In that ground-breaking paper, Mona Baker convincingly argues against the view that translated texts are unworthy of academic enquiry and advocates applying the methods and techniques of corpus linguistics to translated texts. A few years later, Shlesinger (1998) made a similar proposal for interpreting. Since then the field of corpus-based translation and interpreting studies has greatly expanded and matured. Over a quarter of a century later it seems worth while to take stock of the most recent developments. The aim of our study is to carry out a thorough review, both quantitative and qualitative, of recent corpus-based studies of translation and interpreting with a view to describing their key characteristics in terms of data, methods and research foci, identifying potential gaps and suggesting avenues for future research.

The three most recent surveys of translation and interpreting studies (Candel-Mora and Vargas 2013, Zanettin et al. 2015 and van Doorslaer and Gambier 2015) differ from ours in a number of ways, the main one being that all three rely on large bibliographic databases and only analyse the bibliometric records, i.e. the titles, keywords and abstracts, without delving into the full texts of the publications. Our survey is more limited in terms of the number and types of publications, but goes beyond the quantitative, mostly automatic investigation of bibliometric records in order to provide more qualitative insights thanks to an in-depth manual exploration of the full texts of the publications.

There are also differences of scope. The surveys by Zanettin et al. (2015) and van Doorslaer and Gambier (2015) cover much more ground than ours. Their aim is to identify the main subfields and research foci of Translation Studies as a whole, and they therefore provide only limited information on corpus-based studies. However, some interesting findings emerge. Zanettin et al. used the analytic categories in the *Translation Studies Abstracts Online*[1] (TSA) database to identify the most popular subfields of translation and interpreting studies. The results show that the three most popular categories are literary translation, translation theory and intercultural studies. CBTS is only to be found among the next five largest categories, but seems to display an upward trend from 1996 to 2011. As *corpus* and *parallel corpus* also appear as keywords in a corpus of 16,000 abstracts compiled by the authors, they conclude that '[i]n terms of methodologies, the impact of linguistic corpora is noticeable and is a trend that is clearly here to stay' (p. 20). The second survey, by van Doorslaer and Gambier (2015), makes use of the online *Translation Studies Bibliography*[2] (TSB), which currently contains over 30,000 annotated records. An analysis of the authors' academic affiliations provides useful information on the geographical spread of translation and interpreting research. Thanks to the extended list of searchable keywords in the TSB conceptual map, the authors are also able to identify the main topical foci (e.g. literary translation, terminology, teaching) of specific journals and to highlight differences in correlation with the language of publication. Unfortunately, the study fails to reveal any information on corpus-based studies, probably because the keyword analysis is limited to the five most frequent keywords in seven journals.

[1] The TSA content has been merged with the *Translation Studies Bibliography* (TSB).
[2] https://benjamins.com/online/tsb

Unlike the preceding two surveys, Candel-Mora and Vargas-Sierra (2013) focus specifically on CBTS and pursue objectives similar to ours. Their aim is 'to analyze with data the consolidation of corpus methods in translation and to specify which issues are under research and the features that characterize these studies' (p. 317). Unlike our survey, however, their study relies on bibliometric records from two translation databases (*Bibliography of Interpreting and Translation*[3] and *Translation Studies Abstracts Online*). This allows them to provide a wide panorama of the field, based on a large number of publications (389)[4]. As regards the languages represented in the corpora, the survey shows that 40% of the bibliographic records that specify the language refer to a corpus of English or include English in the language pair investigated. The second most represented language, Spanish, falls way behind (13%). The types of corpora used are predominantly parallel corpora (58%), followed by comparable corpora (27%) and a combination of parallel and comparable corpora (15%). The survey also shows that specialized translations are far more numerous (69%) than literary translations (31%). Another interesting finding is that CBTS studies are published in similar proportions as book chapters (45%) and as journal articles (40%). Although the survey provides some useful information on corpus-based translation studies, the bibliometric method on which it is based has its limitations, the main one being that the bibliometric records, because they only contain the title, keywords and words in the abstracts, fail to provide information on many key features of CBTS. For example, only 109 out of 389 records (28%) specify the type of corpus used, and only 11, a mere 3%, refer to the size of the corpus, which considerably reduces the reliability of the conclusions drawn in respect of these aspects. As regards corpus orientation (research, teaching or professional), the authors acknowledge that 'this parameter cannot be interpreted appropriately without carrying out an in-depth study of the publications' (p. 324).

[3] http://dti.ua.es/en/bitra/introduction.html

[4] The authors do not specify the dates of these publications.

Alongside bibliometric analyses, which ‘have the ability to offer factual, quantitatively based, but sometimes also broader views on tendencies in a discipline’ (van Doorslaer and Gambier 2015: 317), we believe there is scope for a survey of corpus-based translation and interpreting studies which relies on manual, in-depth exploration of the actual texts of the publications, thereby allowing for the investigation of qualitative information which is not captured – and indeed in some cases cannot be captured – by an analysis of bibliometric records. Our survey is based on a relatively small number of scientific articles, but as these studies cover the most recent years (2012-2019), they provide a worthwhile snapshot of the latest trends in the field and point the way forward for further research.

The chapter is structured as follows. The first two sections specify the scope of the survey (Section 2) and the method used to extract the survey data (Section 3). The next sections present the results of the survey. Section 4 subcategorizes the corpus-based studies in terms of three main types of corpus orientation, which are analysed in the following sections: methodology- and theory-oriented studies in Section 5, applied studies in Section 6 and empirical studies in Section 7. The dominant category, that of empirical studies, is further explored along three main axes: linguistic focus and translation features, corpus design, and corpus techniques and statistical testing. Section 8 sums up the main findings of the survey and makes some recommendations on desirable developments in the field.

2. Delineating the scope of the survey

In the framework of the present survey, it is essential to establish what qualifies as a bona fide corpus-based study. Our starting point is Baker’s 1993 paper, which sums up the key features of CBTS:

> This paper explores the impact that the availability of corpora is likely to have on the study of translation as an empirical phenomenon. It argues that the techniques and methodology developed in the field of corpus linguistics will have a direct impact on the emerging discipline of translation studies, particularly with respect to its theoretical and descriptive

> branches. The nature of this impact is discussed in some detail and brief reference is made to some of the applications of corpus techniques in the applied branch of the discipline' (p. 233; our underlining).

The first key point is that CBTS is an empirical approach to translation, situated within descriptive translation studies: 'Through the 1970s and beyond, descriptive translation studies (DTS) foregrounded description of what translation was and is, removing from dominance previous approaches that were more concerned with prescribing what translation should be. Corpus-based studies in translation are clearly aligned with the descriptive perspective' (Olohan 2004: 10). For Laviosa (2011: 14), '[t]he strong links forged in those years between Corpus Linguistics and DTS thanks to a set of common concerns stemming from an empirical perspective is (…) one of the keys if not the key to the success story of CTS [corpus-based translation studies]'. Secondly, Baker clearly underlines that CBTS is an offshoot of corpus linguistics, from which it borrows its specific techniques and methods. These involve both 'basic text processing operations' (Baker 1995: 226) such as concordancing and word frequency profiling, and more sophisticated techniques, such as automatic annotation and extraction of keywords, collocations, colligations and word clusters (Zanettin 2012). Finally, Baker underlines the three main objectives of CBTS: to contribute to the theoretical, descriptive and applied branches of translation studies. In the article, Baker further specifies the scope and objectives of CBTS. The term *corpus* is key in CBTS, and Baker insists on its ambiguity in translation studies: 'although the words *corpus* and *corpora* are beginning to figure prominently in the literature on translation, they do not refer to the same kind of corpora that we tend to talk about in linguistics' (1993: 241). In a later article she returns to this issue and specifies what exactly is meant by *corpus* in CBTS: 'any collection of running texts (as opposed to examples/sentences), held in electronic form and analysable automatically or semi-automatically (rather than manually)' (Baker 1995: 225). The way we need to understand the term *corpus* in CBTS is therefore quite different from the way it is regularly used in translation studies, namely to refer to 'fairly small collections of text which are not held in electronic form and which are therefore searched manually' (ibid.). Baker also insists on the key role played by methodology in CBTS, and more particularly by the 'new software tools' and 'new and sophisticated methodologies' (Baker 1993: 248) borrowed from corpus linguistics, which can help counter 'the heavy reliance on introspective methods in translation studies' (ibid.: 240). Large size is also a key characteristic of corpora. Sinclair (1996) lists 'quantity' as a default value of corpora: 'A corpus is assumed to contain a large number of words. The whole point of assembling a corpus is to gather data in quantity'. According to Baker (1993: 237), the fact that translation scholars can now study large numbers of texts of the same type 'is precisely where corpus work comes into its own', as it 'enables the discipline to shed its longstanding obsession with the idea of studying individual instances in isolation (one translation compared to one source text at a time)'.

In the light of this description we decided to limit our survey to translation and interpreting studies that rely on machine-readable corpora, i.e. electronic collections of texts, and are analysed with the help of (semi-)automatic corpus linguistic techniques. The corpora can be monolingual or bilingual/multilingual, comparable or parallel. Size was not selected as a defining criterion as it is a relative notion. In addition, as rightly pointed out by Fernandes (2006: 88), although large size is one of the attributes of corpora, in the context of translation and interpreting studies, corpora are often relatively small and 'the issue of corpus size (…) becomes a relative one in the sense that qualitative aspects sometimes may be more relevant than quantitative ones'.

3. Survey dataset: extraction and general overview

The survey is based on scientific articles written in the years 2012 to 2019 in the following twelve journals: *Across Languages and Cultures, Babel, Interpreting, inTRAlinea, Journal of Specialised Translation, Meta, Perspectives, Target, The Interpreter and Translator Trainer, Journal of Translation and Technical Communication Research* (*trans-kom*), *Translation &*

Interpreting and *Translation and Interpreting Studies.* The selection was made on the basis of two criteria: the journals had to be peer-reviewed and we needed to have direct access to the full texts in our academic environment. The survey is synchronic because the limited period covered (eight years) is not a realistic basis on which to carry out a diachronic study. However, some passing remarks on evolutionary trends will be made where appropriate.

The filtering of corpus-based studies involved the following steps:

- Automatic extraction of the articles written in English[5] that contain the word *corpus* or *corpora* in the title, abstract and/or keywords; full text search if the two words are absent from these sections;
- Manual filtering: rejection of the articles that do not fit our defining criteria, i.e. use of data in electronic format and of corpus analytic methods – from the most basic to the most sophisticated.

As pointed out in Section 2, the presence of *corpus/corpora* is not enough to qualify articles as bona fide corpus-based translation and interpreting studies. As a result, the texts of the 265 articles containing the words *corpus* or *corpora* were scanned in order to ensure that the data were in electronic format and the analysis relied on corpus analytic methods. Table 1.1 gives an overview of the dataset before and after this manual filtering step.

Journal	**Articles in English**	**Articles in English with *corpus/corpora* (unfiltered dataset)**	**Corpus-based articles in English (filtered dataset)**
Across Languages and Cultures	92	34	27
Babel	183	33	16
Interpreting	80	12	8
inTRAlinea	123	26	22

[5] According to Zanettin et al. (2015), 74% of the papers included in TSA were originally written in English. Five of the journals included in our survey publish articles in languages other than English (*Babel*, *inTRAlinea*, *Journal of Specialised Translation*, *Meta* and *trans-kom*).

Journal of Specialised Translation	164	19	12
Meta	132	27	20
Perspectives	303	54	36
Target	128	23	18
The Interpreter and Translator Trainer	142	11	7
trans-kom	42	5	4
Translation & Interpreting	117	10	7
Translation and Interpreting Studies	145	11	9
Total	**1651 (100%)**	**265 (16%)**	**186 (11%)**

Table 1.1: Dataset before and after manual filtering

A comparison of the number of unfiltered (265) and filtered (186) datasets shows that a search that relies solely on the presence of the terms *corpus* or *corpora* generates a non-negligible number of studies (79, i.e. 30%)[6] that do not in fact belong in corpus-based translation and interpreting studies and were therefore excluded from the survey. This somewhat unexpected finding reduces the reliability of surveys such as Candel-Mora and Vargas-Sierra (2013) for CBTS and Liao and Lei (2017) for corpus linguistics that omit the manual filtering step and, as a result, fail to take into account the ambiguity of the word *corpus*. Approximately one third of the excluded studies proved to be based on a paper rather than an electronic format. Others did in fact make use of an electronic corpus (e.g. news articles downloaded from newspaper websites) but relied on purely manual methods to analyse the data, regularly using the 'corpus' as a 'repository of examples' (Tognini-Bonelli 2001: 10) to illustrate one or another phenomenon. Overall, these studies fall within the Descriptive Translation Studies paradigm

[6] This percentage does not include studies where the word *corpus* is used in its strictly literary sense of a collection of writings representing a specific author, topic, genre or period. These studies were excluded without being quantified.

(and are indeed in several instances explicitly described as such by the authors) rather than corpus-based translation and interpreting studies.

Table 1.1 shows that the genuine corpus studies only represent 11% of the total number of articles in English. This result ties in with Zanettin et al.'s (2015: 12) bibliometric survey, which shows that corpus-based studies accounted for c. 7% in 2011. The fact that this percentage is higher than that established by the authors for 1997 (c. 3%), coupled with our own average proportion of 11% for the 2012-2019 period, suggests that corpus-based translation and interpreting studies are experiencing an upward trend that reflects the growth of corpus linguistics studies in general (Liao and Lei 2017: 4). It is important to point out, however, that there are marked differences between the journals. As is clearly apparent from Figure 1.1, two journals display higher proportions: 29% in the case of *Across Languages and Cultures* and 18% in that of *inTRAlinea*, a far cry from the much lower proportions in *Translation and Interpreting Studies* and *Translation & Interpreting* (6%), and *The Interpreter and Translator Trainer* (5%). In other words, some journals seem to be more corpus-oriented than others. These differences cannot be attributed to differences in the journals' overall scope, as not a single descriptive section available on the respective websites contains a reference to corpora, corpus linguistics or corpus-based approaches to translation and interpreting.

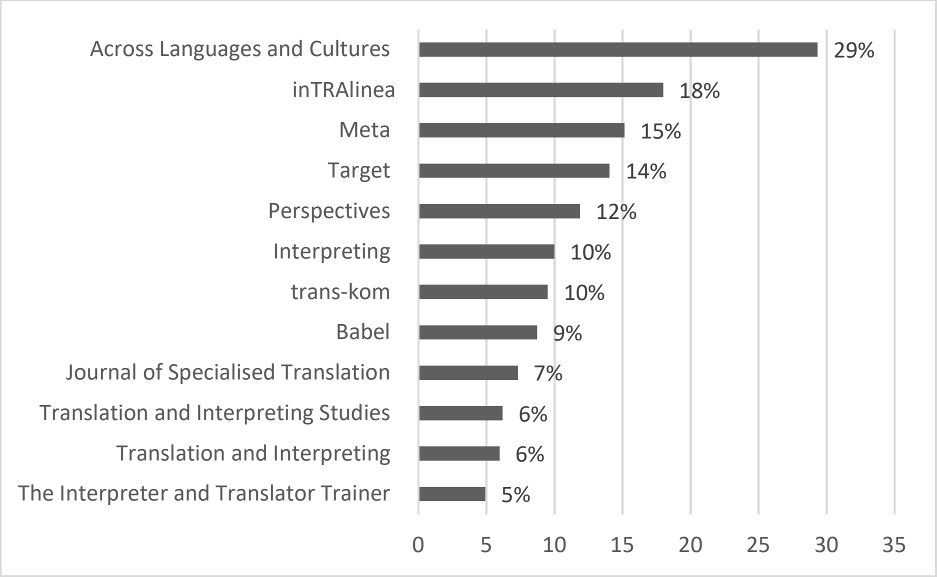

Figure 1.1: Proportion of corpus-based articles per journal

4. Corpus study orientation

We have classified the articles included in the database into three main categories, according to their main research focus and objectives: methodological-theoretical, empirical and applied. This categorization corresponds quite closely to the three branches of translation studies outlined in Holmes (1988[2000]), where a distinction is made between *theoretical*, *descriptive* and *applied* translation studies. For Holmes, descriptive translation studies 'describe[s] the phenomena of translating and translation(s) as they manifest themselves in the world of our experience' (ibid., 176). It is the branch of translation studies where the empirical phenomena under investigation hold centre stage. Theoretical translation studies 'evolve[s] principles, theories, and models which will serve to explain and predict what translating and translations are and will be' (ibid., 178). The third branch of translation studies in Holmes's model, applied translation studies, covers several areas, such as teaching and translation aids (in both translator training and professional practice). In his seminal paper, Holmes does not discuss research methods directly, but he acknowledges the crucial importance of the methodological and meta-theoretical dimension of translation studies, 'concerning itself with problems of what methods and models can best be used in research in the various branches of the discipline (how translation theories, for instance, can be formed for greatest validity, or what analytic methods can best be used to achieve the most objective and meaningful descriptive results)' (ibid., 183).

In the present survey, the empirical category includes corpus studies devoted to specific linguistic phenomena (e.g. grammatical, lexical) and translation features (e.g. explicitation, normalization). The methodological-theoretical category subsumes three main types of contribution: (i) calls for methodological and theoretical advancement, such as proposals for the adoption of methods and theories borrowed from neighbouring disciplines, (ii) literature reviews and overviews, and (iii) descriptions of new corpora and corpus tools for translation and interpreting studies. The applied category covers four major types of corpus application in translation and interpreting studies, namely corpus use in (i) translator and interpreter training, (ii) professional practice (language industry), (iii) translation quality assessment, and (iv) machine translation. It is important to recognize that the three main categories – empirical, methodological-theoretical and applied – are not watertight. For example, although their focus is primarily empirical, a number of studies in the dataset discuss – in more or less detail – the (possible) implications of their descriptive results for theory, methodology or practice. In the same vein, some studies in which methodology and theory hold centre stage include empirical case studies whose purpose is to illustrate the potential for theoretical development or the application of a particular corpus-based method in translation and interpreting studies. Similarly, some applied studies include empirical case studies showcasing practical corpus applications. Despite the relative porousness of the categories, an in-depth analysis of the articles allowed us to assign each study to a single category.

As can be seen in Figure 1.2, empirical studies take up the lion's share, as they account for two thirds of the articles in the dataset, with applied studies and methodological-theoretical contributions lagging far behind (19% and 15% respectively).

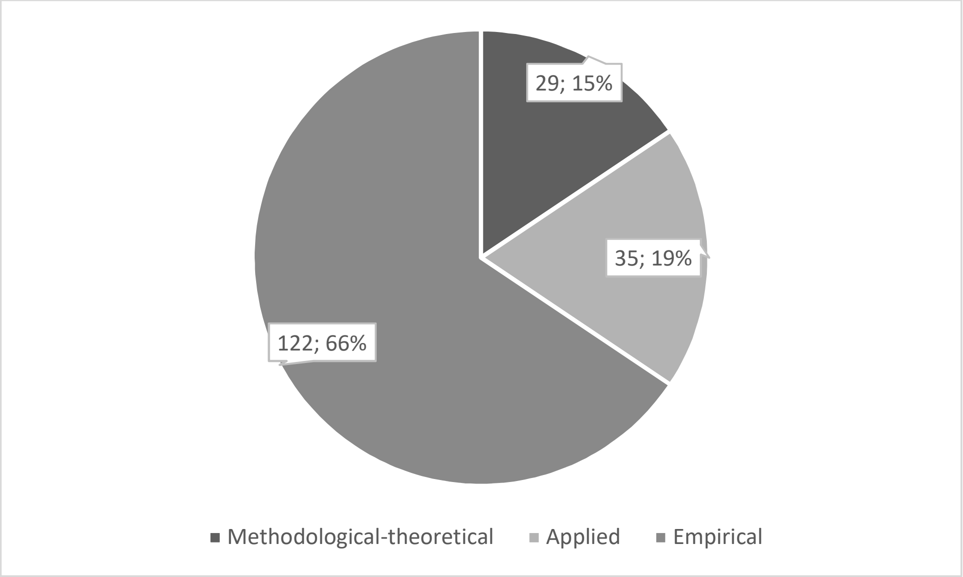

Figure 1.2: Breakdown of empirical, methodological-theoretical and applied studies in the dataset (n=186)

5. Methodology- and theory-oriented studies

Methodological-theoretical contributions represent a sixth of the articles under scrutiny. Even though they are admittedly far less numerous than empirical studies, they testify to the growing maturity of the field in terms of both methodological and theoretical advancement. Methodology- and theory-oriented papers in the dataset are evenly distributed across the three subcategories outlined above (new methods/theories, literature reviews, new corpora/tools). First, we find articles that aim to introduce new methodological and theoretical approaches or concepts, often borrowed from neighbouring disciplines, such as monolingual corpus linguistics or contact linguistics. The methodological aspects concern corpus compilation (e.g. compilation of multimodal corpora), corpus annotation (e.g. annotation of speech acts for discourse analysis), corpus data extraction (e.g. semantic relations in terminology) and data triangulation (e.g. combining corpus and experimental data). Theoretical contributions range from presentations of general frameworks (e.g. cognitive approaches to translation) to discussions of specific constructs (e.g. explicitation). Second, a few articles offer literature reviews of the field or of corpus use in specific areas of translation and interpreting studies. Contributions that are relatively broad in scope focus on some of the evolving contours of corpus-based translation and interpreting studies, for instance in terms of the forms of interlingual translation that are typically investigated (vs the ones that are emerging) or the quantitative methods used in the field. Articles that are more focused in their scope deal with specific areas, such as corpus use in terminology, literary translation or news translation, placing particular emphasis on the added value of cross-fertilization between corpus-based translation and interpreting studies and disciplines such as corpus-assisted discourse studies, digital humanities or stylometry. The third category deals with new corpora and corpus tools for translation and interpreting studies. In our dataset, we mainly find descriptions of interpreting corpora (e.g. signed language interpreting, telephone interpreting) and audio-visual translation (e.g. dubbing), which is in line with the forms of interlingual mediation that have recently entered the field (cf. Defrancq et al. 2015). Some of these corpus descriptions include aspects related to annotation (e.g. annotation of turns in telephone interpreting), but generally speaking, corpus annotation is rather infrequently addressed in the methodology-oriented papers.

6. Applied studies

Around a fifth of the articles published in the selected journals deal with corpus use in various applied areas of translation and interpreting studies, which shows that corpora also have their place in the applied branch of the field. Figure 1.3 shows that the most prominent of these areas is translator and interpreter training. The publications related to translator and interpreter education reflect a rich array of didactic approaches that all rely on the use of electronic corpora in the translation or interpreting classroom (cf. Zanettin et al. 2003, Loock 2016). In other words, these studies show how corpora can be used as translation or interpreting tools. Three quarters of the papers in this category deal with written translation, both general and specialized (e.g. legal or scientific translation), into the trainee translators' L1 or L2. A wide range of corpus types are used: target-language monolingual corpora (whether reference, web or specialized), parallel corpora made up of professional translations, learner translation corpora (i.e. parallel corpora comprised of student translations, whether error-annotated or not), bilingual comparable corpora, and combinations thereof. The main pedagogical objectives of corpus use are related to awareness-raising, decision-making while translating (e.g. to solve translation problems) and revision (editing). Some contributions cover a wide range of phenomena while others have a specific linguistic focus, mostly lexis, phraseology and terminology (e.g. reporting verbs). Audio-visual translation is also represented, though far less than written translation (two papers). The remaining quarter of the articles in this category deal with the use of corpora in interpreter training, whether for simultaneous, consecutive or dialogue interpreting tasks. The focus here is on the use of corpora for interpreting task preparation (e.g. creation of corpus-based bilingual glossaries or term bases) and for materials design (e.g. the use of spoken corpora as source speeches for interpreting exercises). As shown in Figure 1.3, the remaining applied articles account for a quarter of the category and are equally distributed across corpus use in professional practice, quality assessment and machine translation. All three categories are very marginal in the dataset. The use of electronic corpora in the language industry is hardly discussed in our dataset (see Candel-Mora and Vargas-Sierra 2013: 324 for a similar observation) but the few contributions in this area mostly deal with terminology and how corpora can be used to develop terminological resources for professional translators and other language professionals. Our survey also seems to indicate that quality assessment is rarely investigated with corpus methods. Studies on the use of corpora in machine translation are much more widespread, but they are typically published in Natural Language Processing publication outlets and in journals specifically dedicated to machine translation.

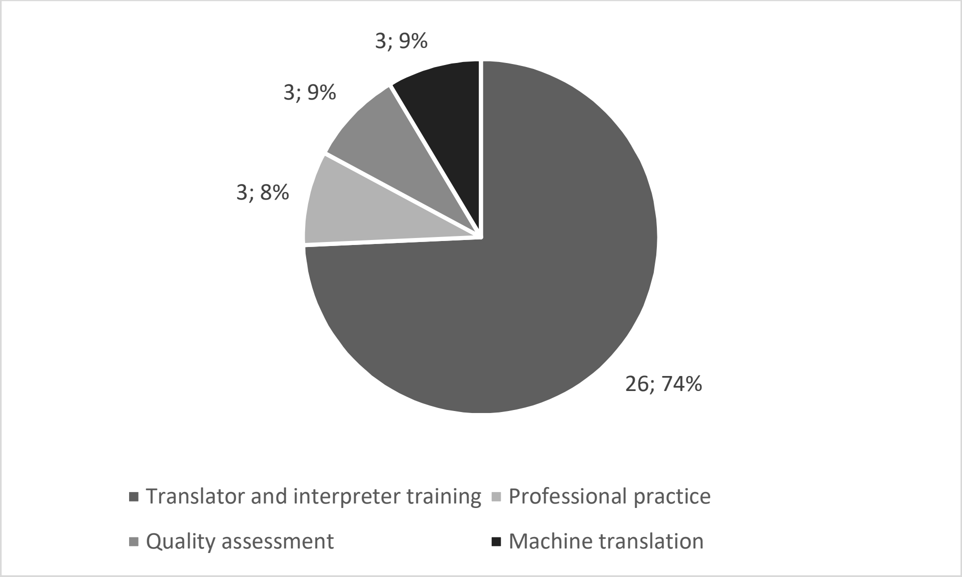

Figure 1.3: Breakdown of applied studies in the dataset (n=35)

7. Empirical studies

The present section offers an overview of the empirical studies included in the dataset. Each empirical study was analysed in terms of its research focus and corpus design, and the corpus techniques and statistical tests used.

7.1 Research focus

A key aspect of research syntheses consists in determining the topical issues that dominate the field under review. To achieve this aim, the 122 empirical studies in our survey data were examined in order to identify their main linguistic focus as well as the attention given to translation features.

7.1.1 Linguistic focus

Empirical studies can be categorized according to the language features they investigate either as the direct focus of the study (e.g. modals) or as a way of assessing the validity of a specific translation feature (explicitation, normalization, etc.). We have grouped these language features into seven broad categories: lexis and terminology (L&T), grammar (G), morphology (M), semantics (S), discourse and pragmatics (D&P), speech-specific (SP) and mixed (MX). Figure 1.4 shows the breakdown of the categories in the 122 empirical studies in our survey.

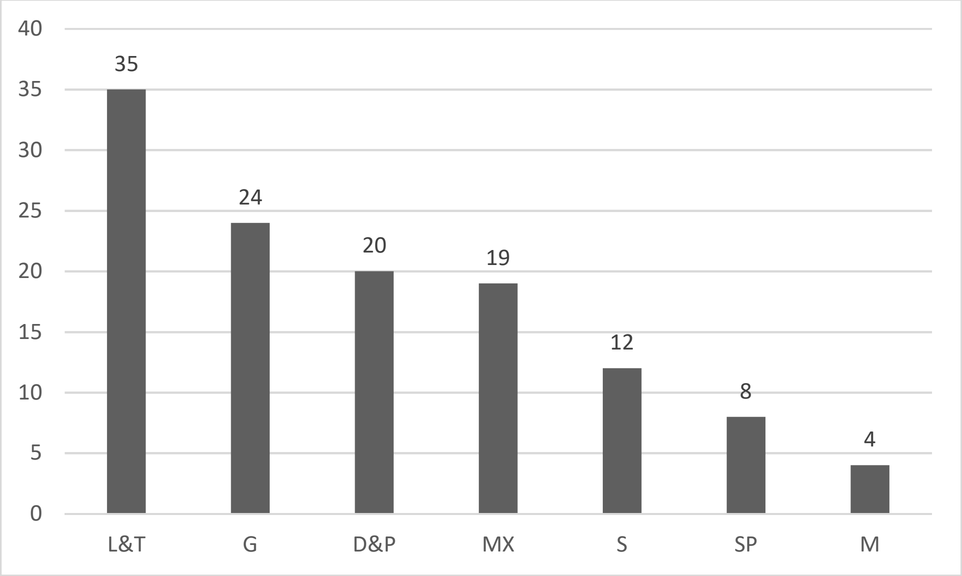

Figure 1.4: Linguistic focus of the empirical studies (n=122)

The most populated category is that of lexis and terminology, which encompasses single words and multi-word expressions as well as single terms and multi-terms. Although phraseology is currently regarded as a category in its own right, we have included it in the L&T category because the dividing line between some of the categories, notably compounds and collocations, is very difficult to draw, especially when it comes to specialized vocabulary. Two factors can explain the dominance of the lexical category: first, lexis in the wide sense has always been at the forefront of translation studies and, second, it is the aspect of language that is the most amenable to corpus techniques, a factor that contributes to the popularity of lexical studies in corpus linguistics, as shown by Gilquin and Gries's (2009: 10) survey. Terms (e.g. business terms) account for 40% of the lexical items investigated; the other studies focus on general vocabulary – either single words (e.g. *between*) or one specific category of words (e.g. phrasal verbs) – or general measures of lexical richness, in particular lexical variation and lexical density.

Grammar, the second most represented category, includes a wide range of grammatical and syntactic phenomena, some of which (passives, modals, nominalizations) recur in the dataset. The D&P category is dominated by discourse-oriented studies, with cohesion, and more particularly the use of connectors, as the main object of investigation. Pragmatics, which is mainly associated with speech, a medium that is in a minority in our dataset compared to writing (see Section 7.2.3), is limited to a handful of studies on (im)politeness. Semantics, which is not easy to approach using corpus techniques, is a relatively minor category limited to the analysis of a few topics such as metaphors and the expression of manner-of-motion. The SP category, which groups studies focused on speech-specific features (pauses, speech rate, hesitations, interactional non-renditions) is also scantily represented in the data. Morphology ranks last, with only four studies focused on derivational affixes. It is interesting that the low frequency of the SP and M categories is not specific to corpus-based translation and interpreting studies as, together with pragmatics, phonology and morphology are among the least frequent categories in Gilquin and Gries's (2009) survey of corpus linguistics. While most studies fall squarely into one well-defined linguistic domain, a sizeable number embrace two or more domains. These studies, which have been grouped into the MX category, aim to provide a general profile of different varieties of translated and/or interpreted language or use a range of language features to operationalize translation features.

7.1.2 Translation features

The identification of 'universal features of translation', i.e. 'features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems' (Baker 1993: 243), was one of the key objectives of early CBTS. At first, the four main features, established by comparing corpora of translated and non-translated texts, were considered to be simplification, explicitation, normalization and levelling out (Baker 1996: 176-7). Since then, other features have been added and the very notion of universality has been called into question, leading researchers to abandon the term ‘translation universal’ in favour of the more realistic ‘translation feature’. Twenty-five years on, it seems worth while to assess the place occupied by this aspect of translation. With a view to identifying the studies that have translation features as their main focus, we scanned the titles and keywords of the studies in our dataset for the following words: *translation universal, translation feature, features of translated language, explicitation, implicitation, normalization, standardization, simplification, levelling out, unique items (hypothesis)* and *convergence*[7]. The results show that translation features remain a strong research strand in current corpus studies, as 29% (35/122) of the studies were extracted on the basis of this criterion. It is important to bear in mind, however, that this percentage does not take into account the many studies that refer to translation features in the analysis (typically, when interpreting their results) but do not highlight them explicitly in the title and/or keywords. Explicitation is by far the most researched feature, either as the sole focus of the study or alongside other features. Normalization/standardization and simplification are also popular, the other features (levelling out/convergence, unique items) trailing far behind. A wide diversity of linguistic phenomena is used to operationalize the translation features. For example, the rate of explicitation is established on the basis of the use of connectors, modals, passives, collocations and manner-of-motion verbs and omission of the conjunction *that* in English. In several instances, translation features are assessed on the basis of a mixture of words, phrases and structures (the MX category) rather than a single linguistic phenomenon.

7.2 Corpus design

[7] While some authors include interference (or source-language influence) in the list of translation features, we follow Baker (1993) who explicitly excludes it from her definition of ‘translation universals’.

This section provides an overview of the corpus designs of the empirical studies under scrutiny, focusing on four aspects: corpus types, corpus size, registers and languages.

7.2.1 Corpus types

We have distinguished between three main categories of corpus type: parallel corpora, monolingual comparable corpora and mixed corpora (combinations of parallel and comparable corpora). As shown in Figure 1.5, parallel corpora (PARA) are much more frequently used than the other corpus types, with more than half of the empirical studies relying on parallel corpus data. They are followed by monolingual comparable corpora (CMONO, 25%) and mixed corpora (MIX, 16%). Other corpus types are rarely used (6%). The clear dominance of parallel corpora over monolingual comparable corpora sheds light on the central position of source text-target text (ST-TT) comparisons in the field. While this is in sharp contrast with Baker's (1995: 233) programmatic call for a shift away from ST-TT comparisons to comparisons of translation with original text production, it reflects a concern voiced quite early in the field, namely that translation products cannot be fully understood if they are cut off from their ST (see e.g. Stewart 2000 and Kenny 2005 on the 'target-orientedness' of Baker's and other early corpus-based work). Interestingly, the figures presented here are very similar to the ones obtained by Candel-Mora and Vargas-Sierra (2013) in their bibliometric-based survey covering approximately 15 years, which seems to indicate that parallel corpora have dominated the field for quite some time. Back in 2005, however, Kenny (2005: 155) insisted that the use of parallel corpora had been strikingly limited in the first decade of CBTS research. A larger-scale survey would be needed to track this development more precisely.

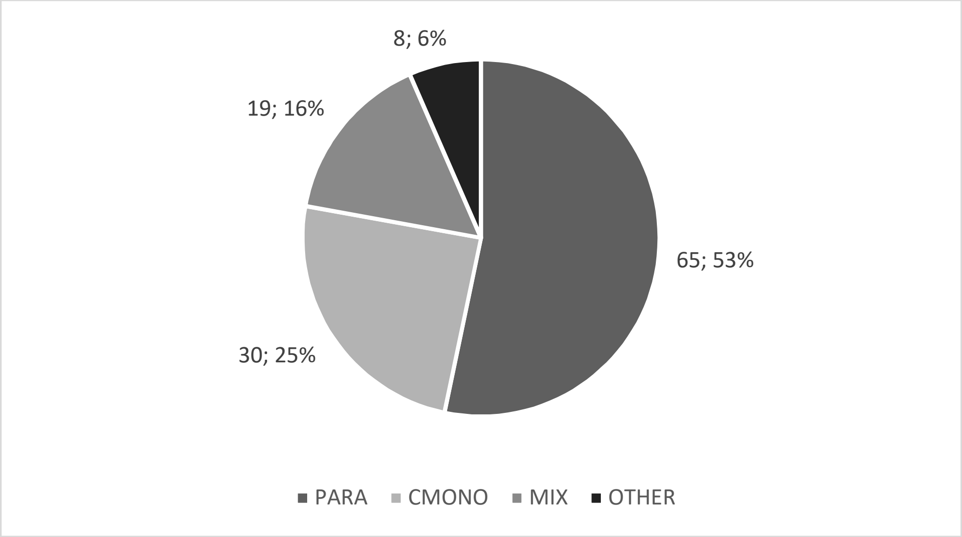

Figure 1.5: Corpus types in empirical studies (n=122)

Parallel studies are typically monodirectional and restricted to one language pair (e.g. English>Spanish). There are also a few monodirectional studies that examine different language pairs (e.g. French>English and French>Dutch) or make use of several same-language versions of the STs, such as unedited and edited versions of the same texts and different translations of the same novel. We find very few cases of bidirectional studies, i.e. studies where the two translation or interpreting directions are analysed (e.g. English>Italian and Italian>English). This is particularly regrettable, as the investigation of the two directions makes it possible to go some way towards disentangling source-language (SL) influence from more general translation features. Bidirectional parallel corpora have shown their worth in contrastive linguistics, where it is common practice to examine the two directions in order to identify cross-linguistic correspondences. In the empirical studies of our dataset, researchers typically rely on parallel corpora to examine the procedures used to translate given ST items or structures (e.g. culture-specific items, passives) and the explicitation/implicitation of certain ST phenomena, such as connectors (explicitation is the only translation feature to often be approached from a parallel perspective).

Monolingual comparable corpora, which account for a quarter of the empirical studies under investigation, are often made up of two subcorpora, namely translated/interpreted language and non-translated/non-interpreted (i.e. original) language. In c. 75% of these studies, the texts are translated (or interpreted) from a single source language (e.g. Arabic translated from English). Less commonly, the texts/speeches have been translated/interpreted from several SLs, which makes it possible to study the effect of SL influence. Monolingual comparable corpora are mostly used to examine translation features (e.g. normalization, simplification, increased explicitness, unique items), in line with Baker's research agenda for the field.

Around a sixth of the empirical studies rely on a combination of parallel and comparable corpora. The ways of combining the two types of corpus are manifold, depending for instance on whether it is the parallel or the comparable perspective that holds centre stage. The most common approach in the dataset at hand is the combined use of a monodirectional parallel corpus and a comparable corpus of target-language original texts. Typically, the former is the core of the study, while the latter is used to check whether the trends identified in the TTs of the parallel corpus diverge from the ones found in a comparable set of original texts (e.g. the frequency of the passive voice).

Other corpus types are far less frequent. They include multilingual comparable corpora (representing two or more original languages), which are analysed with reference to translation-related objectives (e.g. cross-linguistic register description), and monolingual corpora (e.g. audio-description). More generally, irrespective of corpus types, we see that the corpora used in the empirical studies mainly contain professional L1 translations/interpretations (exceptions include learner translation corpora and corpora of Chinese>English retour interpreting).

7.2.2 Corpus size

Unsurprisingly, a close inspection of corpus size reveals that the parallel and comparable corpora used in the empirical studies are rather small by today's standards in mainstream corpus linguistics. For monolingual comparable studies, we find that approximately half of the corpora are smaller than 1 million tokens in total (all subcorpora considered), while the other half are larger than 1 million tokens. Parallel corpora tend to be much smaller in size. The survey shows that c. 40% of the parallel corpora contain less than 100,000 ST tokens and another 40% between 100,000 and 1 million ST tokens. This is in part due to parallel corpora of interpreting, which are smaller on account of the many hurdles inherent in transcribing speech (Bernardini et al. 2018).

The relatively small size of the corpora used is not necessarily problematic, as some analyses do not require large corpora, for example because they focus on high-frequency phenomena or involve the manual coding of numerous variables. In this regard, an interesting methodological solution, found in a few empirical studies in our dataset, is the use of large general reference corpora to compensate for the relatively small size of the corpora used. For example, phraseological units, such as collocations, can be identified in translated texts on the basis of the statistical association scores they display in large reference corpora.

It is important to point out that size figures could not be retrieved for all the empirical studies. In 19% of the cases, corpus size is not mentioned or is not provided in tokens (but rather in number of texts or length in minutes). It seems that De Sutter et al.'s (2012: 137) methodological call for research papers in the field to 'provide a meticulous overview of the corpus materials used' has not yet been fully heeded.

7.2.3 Modality and registers

Three quarters of the empirical studies examine translation (whether written or audio-visual). Interpreting accounts for a little over a fifth of the empirical studies. A handful of studies are intermodal: they compare written translation and simultaneous interpreting. These figures point to a noticeable breakthrough by corpus-based interpreting studies in recent years (cf. Russo et al. 2018). Interpreting studies in the dataset examine a wide range of interpreting contexts, such as government press conferences, parliamentary debates and court proceedings. Even though most studies deal with simultaneous interpreting, other forms of interpreting are investigated as well, such as dialogue interpreting. The spoken corpora used in the studies are mostly comprised of transcripts of spoken data, following various transcription conventions, and, less frequently, verbatim reports of parliamentary debates.

The survey further shows that four main types of written register are investigated: literature (LIT), specialized registers (SPEC), audio-visual registers (AVT) and news (NEWS). Literature and specialized registers rank first, followed by audio-visual registers and news (see Figure 1.6). The sizeable proportion of empirical studies devoted to literary translation is in line with Zanettin et al. (2015), who found that literary translation is one of the top three most-researched topics in translation and interpreting studies. As can be seen in Figure 1.6, multi-register studies (MULTI) are also widespread in the dataset.

While the LIT category focusses almost exclusively on a single register, namely novels, SPEC is very fragmented, and covers registers as diverse as legal texts (e.g. treaties), reports emanating from international institutions and popular science articles. The AVT category is also quite diversified, with studies devoted to subtitling, dubbing, voice-over and audio-description. Movies and sitcoms figure prominently in the dataset, whatever the type of AVT. AVT corpora mainly contain textual data, but we also find some studies relying on multimodal and multimedia AVT corpora. Surprisingly, news translation is not frequently investigated on its own in our dataset, despite the wide availability of translated news items nowadays. MULTI studies, though quite numerous, are rather difficult to characterize. Some of them offer cross-register analyses. These analyses, which are based on register-stratified corpora such as the

Dutch Parallel Corpus (Macken et al. 2011) and P-ACTRES (Izquierdo et al. 2008), reflect the recent interest in register variation. It is important to note, however, that not all MULTI studies analyse registers contrastively. Rather, in such studies, registers are simply combined (e.g. to make up for lack of sufficient data) and studied as a unified whole.

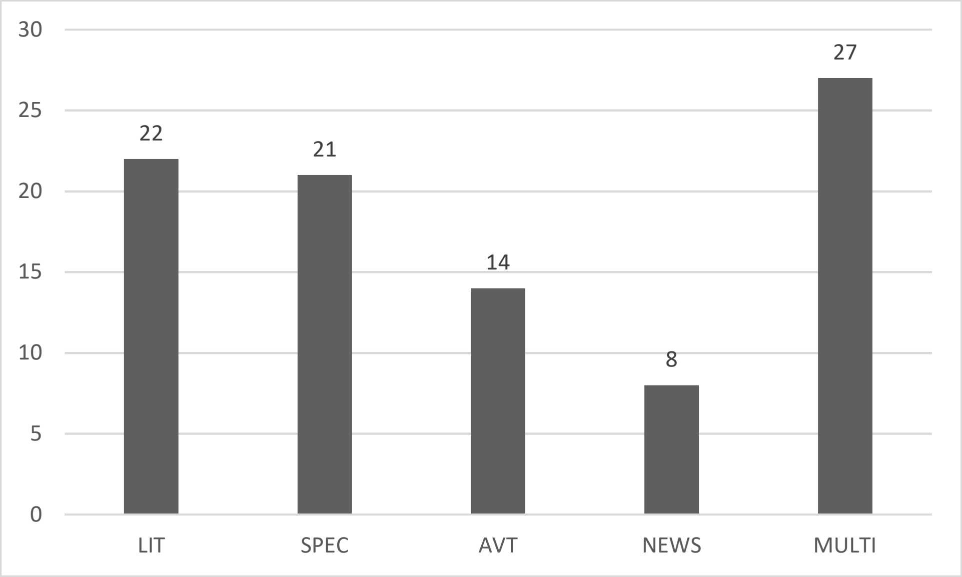

Figure 1.6: Written registers in empirical studies (n=92)

7.2.4 Languages

A total of twenty-three languages are investigated in the empirical studies, of which two thirds are European languages (exceptions include Arabic, Hebrew, Russian and Thai). The survey reveals a clear English-centric perspective, as 72% of the studies deal with English, either on its own or in combination with other languages. In this respect, we agree with Vandevoorde and De Sutter (2019: 1) that 'this hegemony of English raises fundamental issues about the nature and relevance of research questions and theoretical concepts, the stability of research findings and the appropriateness of methodologies that are primarily tailored towards the investigation of the English language'. English is followed, far behind, by French, Spanish, Dutch, Chinese, Italian and German, each of which accounts for between 18% and 11% of the studies. As can be seen, Chinese is the only non-European language that features in this top list (it is mainly examined in interpreting studies in our dataset).

7.3. Corpus techniques and statistical testing

7.3.1 Corpus techniques

As pointed out in Section 2, a study only qualifies as a bona fide corpus study if it relies on the automated techniques developed within the framework of corpus linguistics. This part of the survey, which required a minute scanning of the 122 studies, proved to be particularly arduous as the information was often incomplete and tended to be scattered rather than included in a separate section devoted to data and methodology. Several studies merely reported that the occurrences of the phenomenon under scrutiny were 'extracted' or 'identified' without providing any information on the way they were extracted and/or processed. The in-depth exploration of the texts proved to be well worth the effort, however, as it allowed some interesting trends to emerge. The main finding is that the majority of the studies (c. 60%) rely solely on the 'basic text processing operations' that were described by Baker (1995: 226) in the early days of CBTS, over twenty-five years ago, i.e. frequency and concordancing. In some studies, the only aspect that is computed automatically is the number of words in the corpora used, the bulk of the study being carried out manually. Although this approach makes minimal use of the electronic nature of the data, it is still quite valuable in that it makes it possible to compare the frequency of linguistic phenomena across corpora (e.g. in translated and non-translated language), using both raw and relative frequencies. Most studies, however, go one step further and make use of the word list and concordancing functionalities offered by text analysis software. Word lists are particularly useful as they provide the frequency of all words in the corpus data used and make it possible to compute lexical variation indices (type/token ratio and standardized type/token ratio) automatically. Concordancing lives up to its reputation as 'the corpus analyst's stock-in-trade' (Baker 1995: 226), the two most popular programs being the monolingual tools *WordSmith Tools* (WST) (Scott 2016) and *AntConc* (Anthony 2019). Bilingual concordancers such as *ParaConc* (Barlow 2008) are much less frequently used. In many cases, however, the authors analyze concordance lines but provide no indication of the program used. Of the 45 studies that contain an explicit reference to one of the above-mentioned three programs, a deplorably high number (15) fail to include a bibliographic reference.

Some 40% of the studies make use of more advanced techniques. We have classified as advanced all the techniques that go beyond word-form-based extraction and concordancing. Two main categories of technique emerge: automatic annotation and automatic extraction of keywords and phraseological units. The most popular types of annotation in the survey are lemmatization and part-of-speech (POS) tagging. The first relieves the researcher of the burden of extracting the inflected forms of one and the same lemma and is a necessary step for computing the lexical density of texts. The second allows researchers to carry out extractions focused on a whole word category (e.g. modal verbs) and relieves them of another burden, that of disambiguating homonyms (e.g. the noun *can* vs the auxiliary *can*). The automatic extraction of keywords and phraseological units relies on frequency, co-occurrence and recurrence indices and can be performed automatically by programs such as *WST* and *AntConc.* Several studies in the survey use this method to extract collocates, lexical bundles, keywords and key clusters. A number of studies make use of the *Corpus Workbench*[8] whose central component is a powerful query processor that makes it possible to query large corpora with linguistic annotations. Surprisingly, only two studies make use of *Sketch Engine*, which contains very powerful functionalities for translation research and allows translation scholars to upload their own corpora (Kilgarriff et al. 2014). Several studies focused on speech make use of *EXMARaLDA*

[8] http://cwb.sourceforge.net/index.php

(Schmidt and Wörner 2009), a system for working with oral corpora which includes a transcription and annotation tool and a query and analysis tool. While the majority of the studies rely on independent software programs, a non-negligible number base their analysis on corpora such as P-ACTRES (Izquierdo et al. 2008), which come with their own interface supporting basic and complex queries on word forms, lemmas, POS tags and phrases.

The results show that the majority of the corpus-based translation and interpreting studies in our survey do not exploit the full potential offered by the electronic nature of the corpus. They tend to rely on a limited set of basic corpus techniques and thereby fail to display one of the key characteristics of corpus-based studies, i.e. the fact that they 'make extensive use of computers for analysis' (our underlining) (Biber et al. 1998: 4). Resources exist for researchers who would like to exploit corpus techniques more intensively. In particular, Zanettin's (2012) and Mikhailov & Cooper's (2016) volumes are excellent sources of inspiration. There are signs that the situation is changing, however. A breakdown of the two approaches – simple vs advanced – per year shows that 55% of the studies using advanced techniques were published in the last two years covered by the survey (2018 and 2019), while the percentage of those relying on more basic methods is only half as great (27%). This said, it should be stressed that sophisticated corpus techniques are not required and indeed are not even practicable for many types of study, particularly those focused on aspects of language – semantic, functional or cultural – that are very hard to handle automatically.

7.3.2 Statistical testing

In addition to corpus techniques, we examined whether statistical tests were used in the 122 empirical studies. To do so, we relied on the distinction between *descriptive* statistics, i.e. statistics that describe and summarize datasets, such as frequency counts, and *inferential* statistics, i.e. statistics that make it possible to infer whether a trend observed in a dataset is representative of the whole population sampled. We found that 55% of the papers rely on descriptive statistics only (mostly in the form of relative frequencies), without recourse to inferential statistics (see Figure 1.7). The other studies rely on inferential statistics. Among those, we find that a majority of studies make use of monofactorial tests (whether parametric or non-parametric), typically based on contingency tables or mean rank orders. Examples of such tests include the chi-squared test, the t-test, ANOVA (Analysis of Variance), the Mann-Whitney U test and Pearson's correlation. More elaborate statistical methods, namely multivariate exploratory techniques (e.g. correspondence analysis) and multivariate tests (e.g. regression modelling, inference trees and random forests) are used in a small number of empirical studies, most of them published between 2017 and 2019. While the use of elaborate statistical testing and advanced quantitative methods is a most welcome development in corpus-based translation and interpreting studies (cf. De Sutter et al. 2012), it is also important to warn against an excessive drift in focus from linguistic description to statistical analysis. In their survey, Larsson et al. (2020) find that the steady increase in the use of advanced statistical methods in corpus linguistics is coupled with a decreased focus on linguistic description. They therefore advocate striking a balance between these two central aspects of corpus work.

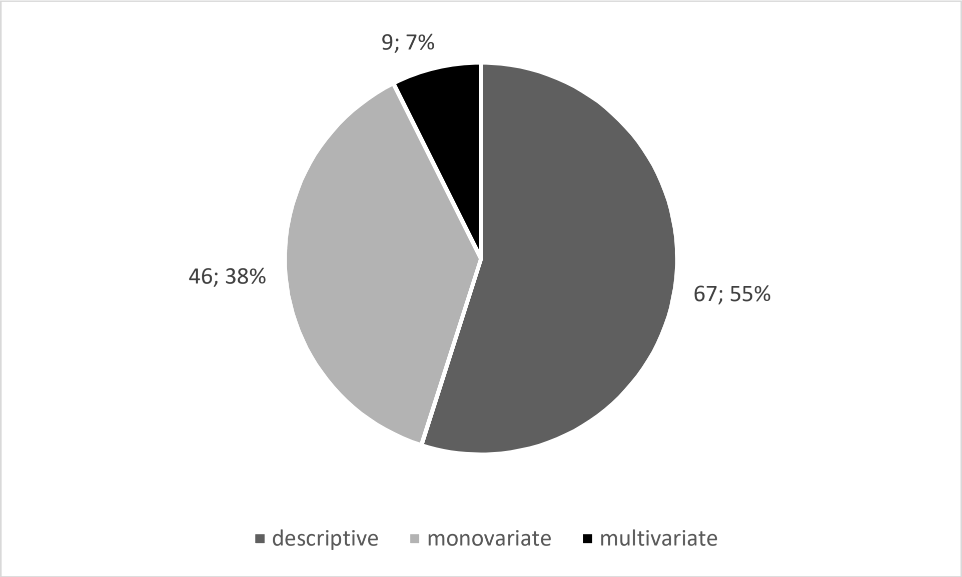

Figure 1.7: Use of statistics in empirical studies (n=122)

8. Conclusion and outlook

Research surveys are ideal instruments with which to take stock of academic fields with a view to identifying current and emerging trends, assessing both strengths and weaknesses, and to suggest directions for future developments. Surveys come in many shapes and forms. One popular type relies on bibliometric data available in online bibliographies, in particular titles, abstracts and keywords. The advantage of this method is that it gives access to a wide range of factual information (number of publications per year, types of publication, range of publication languages, etc.) from a large number of studies published in several formats (books, book chapters, journal articles, proceedings volumes) in a wide range of languages. The disadvantage is that the information that can be extracted from abstracts, titles and keywords is too limited to provide in-depth qualitative insights into the field under scrutiny. The method we decided to use for our survey of corpus-based translation and interpreting studies relies on a minute exploration of the full texts of 186 recent journal articles written in English[9]. It is a very time-consuming method, which precludes investigating a very large dataset. However, as we hope the results of our survey demonstrate, it compensates for this weakness by providing a rich picture of theoretical, methodological and descriptive aspects of the current status of the field. The two approaches are therefore clearly complementary.

The data extraction stage of the survey showed that journal articles that meet the requirements of corpus studies (in terms of data type and corpus techniques) account for 11% of our initial dataset. The breakdown per journal, however, showed that some journals were more corpus-

[9] We recognize that the restriction to studies written in English is a major limitation. It stems from our method of analysis which requires careful reading of the full texts, a task that we cannot undertake in languages we do not master. It should in no way be taken as a lack of recognition of the value of articles written in other languages. Researchers using bibliometric measures can include publications in many languages because they only rely on the abstracts which are written in English.

oriented than others. In addition, the survey brought out a number of key trends in present-day corpus-based studies, testifying to recent developments in the field while also highlighting areas where progress has been relatively modest. First, an analysis of the overall corpus orientation of the studies into three main categories – empirical, methodological-theoretical and applied – showed that empirical studies accounted for two thirds of the studies. In view of the descriptive, product-oriented slant of corpus linguistics, this is not particularly surprising. What came as something of a surprise, however, and can be seen to testify to the growing maturity of the field, is that one third of the studies went beyond description to tackle methodological and theoretical aspects and concrete applications. Second, a detailed scanning of the linguistic focus of each empirical study showed that the dominant category was that of lexis and terminology, followed by grammar, discourse and pragmatics, together with a mixed category comprising more than one linguistic domain. Semantic, speech-related and morphological features turned out to be less popular. Translation features (in particular, explicitation) proved to remain a popular subject of investigation, in line with Baker's research agenda. Third, the analysis of the corpus designs of the empirical studies showed that parallel corpora are used twice as frequently as monolingual comparable corpora, contrary to Baker's (1995) call to move away from ST-TT comparisons. Corpora used in the field were found to represent a wide range of written and spoken registers, with a clear overrepresentation of English (either as a source or target language). Fourth, methodology- and theory-oriented studies proved to be quite diverse, ranging from descriptions of new corpora, literature reviews and calls for the use of more advanced quantitative methods to the application of particular theoretical constructs or models, fostering cross-fertilization with neighbouring disciplines. Finally, applied studies appeared to be mostly geared towards corpus use in translator and interpreter training, while other applied areas, such as corpus use in professional practice or translation quality assessment, were found to be rarely explored.

One of the survey's most important findings concerns the use of corpus techniques and statistics. The analysis showed that the majority of the empirical studies relied on fairly basic techniques (frequency, concordancing), which were promoted in Baker's early papers. More advanced techniques were found to be less frequently used, the dominant types being automatic lemmatization and POS-tagging and techniques to extract keywords and phraseological units automatically. The survey also revealed that most studies rely on simple descriptive statistics (such as relative frequencies) or monovariate inferential statistics, although advanced corpus techniques and elaborate statistical testing have recently started to gain momentum.

The picture drawn by our survey is only partial as it is limited to journal articles written in English and therefore leaves out many relevant publications written in other languages and published in other formats. In spite of these limitations, the study offers a useful survey of the field and allows us to formulate a few forward-looking suggestions. First, there is a need to build new, large corpora for translation and interpreting studies, especially bidirectional parallel corpora. As things stand, the field tends to rely on small ad hoc corpora, very few of which are available to the research community. The new corpora should comprise several registers and involve many languages so as to curb the current dominance of English. There are promising initiatives in this direction, such as the TransBank project[10]. Second, care should be taken to provide a detailed description of corpus data and methodology – corpus type, corpus size, data extraction, selection and annotation, etc. – and to ensure that the information is grouped in one dedicated section rather than scattered across various sections. Third, future studies should aim to exploit the full potential of corpus techniques rather than limiting themselves to frequency profiling and concordancing. In addition, although this is admittedly not in the hands of researchers, the field of corpus-based translation and interpreting studies could be greatly

[10] https://transbank.info/

boosted if translation and interpreting journals were to give more visibility to corpus approaches in the range of topics listed in the description of the scope of the journal.

Corpus-based translation and interpreting studies is still a relatively young research field. It is therefore only natural that some aspects of it have not yet attained full maturity. However, the fact that activity is thriving on all fronts – empirical, theoretical, methodological and applied – is a strong sign that the field will continue to progress unabatedly in the future.

Acknowledgements

We would like to thank Gert De Sutter for his valuable feedback on the first draft of this chapter. Any remaining shortcomings are, however, our own. Thanks also go to Thomas Simon for his help with the extraction of a preliminary version of the dataset.

References

Anthony, L. (2019), *AntConc*, Tokyo: Waseda University.

Baker, M. (1993), 'Corpus Linguistics and Translation Studies. Implications and Applications', in M. Baker, G. Francis and E. Tognini-Bonelli (eds), *Text and Technology: In Honour of John Sinclair*, 233–50, Amsterdam/Philadelphia: Benjamins.

Baker, M. (1995), 'Corpora in Translation Studies: An Overview and Some Suggestions for Future Research', *Target*, 7(2): 223–43.

Baker, M. (1996), 'Corpus-based Translation Studies: The Challenges that Lie Ahead', in H. Somers (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, 175–86, Amsterdam: Benjamins.

Barlow, M. (2008), 'Parallel Texts and Corpus-Based Contrastive Analysis', in M. de los Ángeles Gómez González, J. L. Mackenzie and E. M. González Álvarez (eds), *Current Trends* *in Contrastive Linguistics. Functional and Cognitive Perspectives*, 101–21, Amsterdam: Benjamins.

Bernardini, S., Ferraresi, A., Russo, M., Collard, C. and B. Defrancq (2018), 'Building interpreting and intermodal corpora: A How-to for a formidable task', in M. Russo, C. Bendazzoli and B. Defrancq (eds), *Making Way in Corpus-Based Interpreting Studies*, 21–42, Springer.

Biber, D., Conrad, S. and R. Reppen (1998), *Corpus Linguistics. Investigating Language Structure and Use*, Cambridge: Cambridge University Press.

Candel-Mora, M.A. and C. Vargas-Sierra (2013), 'An Analysis of Research Production in Corpus Linguistics Applied to Translation', *Procedia*, 95: 317–24.

Defrancq, B., De Clerck, B. and G. De Sutter (2015), 'Corpus-based translation studies: Across genres, methods and disciplines'. *Across Languages and Cultures*, 16 (2): 157–62.

De Sutter, G., Goethals, P., Leuschner, T. and S. Vandepitte (2012), 'Towards methodologically more rigorous corpus-based translation studies', *Across Languages and Cultures*, 13 (2): 137–43.

Fernandes, L. (2006), 'Corpora in Translation Studies: Revisiting Baker's Typology', *Fragmentos*, 30: 87–95.

Gilquin, G. and S. Th. Gries (2009). 'Corpora and experimental methods: A state-of-the-art review', *Corpus Linguistics and Linguistic Theory*, 5 (1): 1–26.

Holmes, J. S. ([1988]2000), 'The name and nature of translation studies', in L. Venuti (ed.), *The Translation Studies Reader*, 180–92, London/New York: Routledge.

Izquierdo, M., Hofland, K. and Ø. Reigem (2008), 'The ACTRES Parallel Corpus: an English-Spanish Translation Corpus', *Corpora*, 3 (3): 1–41.

Kenny, D. (2005), ‘Parallel corpora and translation studies: old questions, new perspectives? Reporting that in Gepcolt: a case study’, in G. Barnbrook, P. Danielsson and M. Mahlberg (eds), *Meaningful texts: the extraction of semantic information from monolingual and multilingual corpora*, 154–65, London/New York: Continuum.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and V. Suchomel (2014), ‘The Sketch Engine: ten years on’, *Lexicography*, 1: 7–36.

Larsson, T., Egbert, J. and D. Biber (2020), ‘Do corpus linguists focus on statistics at the expense of linguistic analysis? A ten-year perspective’, conference presentation given at ICAME 41, Heidelberg, Germany [online], 20-24 May, 2020.

Laviosa, S. (2011), ‘Corpus-based translation studies: Where does it come from? Where is it going?’, in A. Kruger, K. Wallmach and J. Munday (eds), *Corpus-Based Translation Studies: Research and* Applications, 13–32, London/New York: Bloomsbury.

Liao, S. and L. Lei (2017), ‘What We Talk about When We Talk about Corpus: A Bibliometric Analysis of Corpus-related Research in Linguistics (2000-2015)’, *Glottometrics*, 38: 1–20.

Loock, R. (2016), *La traductologie de corpus*. Villeneuve d'Ascq: Presses universitaires du Septentrion.

Macken, L., De Clercq, O. and H. Paulussen (2011), ‘Dutch Parallel Corpus: A Balanced Copyright-Cleared Parallel Corpus’, *Meta*, 56 (2): 374–90.

Mikhailov, M. and R. Cooper (2016), *Corpus Linguistics for Translation and Contrastive Studies. A guide for research*, London/New York: Routledge.

Olohan, M. (2004), *Introducing Corpora in Translation Studies*, London/New York: Routledge.

Russo, M., C. Bendazzoli and B. Defrancq, eds (2018), *Making Way in Corpus-Based Interpreting Studies*, Springer.

Schmidt, T. and K. Wörner (2009), 'EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research', *Pragmatics*, 19 (4): 565–82.

Scott, M. (2016), *WordSmith Tools version 7*, Stroud: Lexical Analysis Software.

Shlesinger, M. (1998), 'Corpus-based interpreting studies as an offshoot of corpus-based translation studies', *Meta*, 43 (4): 486–93.

Sinclair, J. (1996), *EAGLES. Preliminary recommendations on Corpus Typology*. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Stewart, D. (2000), 'Poor relations and black sheep in Translation Studies', *Target*, 12(2): 205–28.

Tognini-Bonelli, E. (2001), *Corpus Linguistics at Work*, Amsterdam/Atlanta: Benjamins.

Vandevoorde, L. and G. De Sutter (2019), 'Empirical translation studies in a monolinguistic world: theoretical and methodological challenges' workshop description, EST Congress 2019, Stellenbosch University, South Africa, 9-13 September, 2019.

van Doorslaer, L. and Y. Gambier (2015), 'Measuring Relationships in Translation Studies. On Affiliations and Keyword Frequencies in the Translation Studies Bibliography', *Perspectives: Studies in Translatology*, 23 (2): 305–19.

Zanettin, F. (2012), *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*, Manchester: St. Jerome Publishing.

Zanettin, F., Bernardini, S. and D. Stewart (2003), *Corpora in Translator Education*, London: Routledge.

Zanettin, F., Saldanha, G. and S.-A. Harding (2015), ‘Sketching Landscapes in Translation Studies: A Bibliographic Study’, *Perspectives: Studies in Translatology*, DOI:10.1080/0907676X.2015.1010551