JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Waller, S., Ellen Kerans, M. & Maher, A. (2008). Acquiring or enhancing a translation specialism: the monolingual corpus-guided approach. *The Journal of Specialised Translation, 10*, 56-75.
https://doi.org/10.26034/cm.jostrans.2008.661

This article is publish under a *Creative Commons Attribution 4.0 International* (CC BY):
https://creativecommons.org/licenses/by/4.0

© Stephen Waller, Mary Ellen Kerans, Ailish Maher, 2008

Acquiring or enhancing a translation specialism: the monolingual corpus-guided approach

Ailish Maher, Stephen Waller and Mary Ellen Kerans
Freelance translators/editors, Barcelona, Spain

ABSTRACT

Translators and editors who work in a specialised field—a particular branch of medicine, technology or finance, for instance—may find it difficult to acquire (or enhance) their domain-specific knowledge other than by learning as they go or going back to college. Both strategies can be slow and costly. Our paper describes a faster, more economical way to climb the specialist learning ladder, namely a corpus-guided approach to translating, revising and editing. We describe two tools for analysing a corpus of model texts: on the one hand, a user-friendly concordancer with an intuitive interface; on the other, an equally easy-to-use desktop-based indexer. Finally, we propose an approach to the issue of corpus size (sampling adequacy) that provides a practical solution for the working translator: we recommend creating a carefully chosen, cleaned text collection that functions as a reliable substrate corpus for language pattern guidance and adding to it an ad-hoc 'quick and dirty' corpus to further narrow the topic focus as needed.

KEYWORDS

Corpus, concordancer, translation, editing, corpus-guided translation.

1. Introduction

In the Mediterranean translation market, in which our experience is rooted, higher rates and better working conditions are commanded by specialist translators, editors-revisers of specialist texts, and revisers of specialist translations whose products answer to a very high standard. For translators working in fields in which they lack linguistic confidence, it can be difficult to acquire specialist knowledge other than by learning on the job or going back to college. Novice translators may wonder where to invest their efforts and which specialism might turn out to be their best choice 10, 15 or 20 years down the road. A trial-and-error process is painfully slow and typically results in uneven quality. As for training, there is a shortage of specialist courses, particularly of online short courses that would suit working translators.

A viable alternative is the corpus-guided approach, which consists of systematically collecting target-language texts in the same genre and knowledge area as the source text in order to create a corpus that can be mined using one or several of the software tools available for text analysis.

We have seen that educators of translators are increasingly calling for training in this approach to problem solving (López-Rodríguez and Tercedor-Sánchez, 2008; Wilkinson, 2005a; Varantola, 2003) and we applaud their effort. However, in our experience, relatively few language

service providers (certainly not those working in specialist fields) have formal translation training. We have therefore become involved in continuing education for working translators who come from a variety of backgrounds and who stand to gain a great deal from learning how to design and exploit corpora. Our approach starts from a working translator's point of view rather than an academic one, although our perspective draws heavily on the principles of the late John Sinclair (1991). Although corpus analysis cannot be said to be well established in translation outside of academic circles, it is used widely in applied linguistics (see Hunston, 2002, for an overview and advice on applications), terminology research (for instance, Zweigenbaum, 2003), and in contrastive genre analysis of comparable corpora of relevance to translators (e.g., Williams, 2005; Moreno, 1997); it also underpins various approaches to specialist language teaching (Swales, 1990). The interest of these academics is largely focused on exploring language use in an academic sense and in studying processes and products. The time-strapped working translator or editor, however, is simply interested in emulating 'good' writing in the target language and genre —and this encompasses a wide range of issues such as terminology, word choice, grammar, register and style. Should *data* be used in the singular or plural in the computer field in comparison to other fields?[1] How are forms of *pathology* used in different medical sub-specialisms? What, if any, are the differences in the use of *face value*, *par value* and *nominal value* in a financial context?

Our corpus-guided approach to translation is monolingual: although it is the source text which presents us with a problem (terminology, vocabulary choices, phrasing, sentence patterns, etc), we look for solutions in 'good models for the target text'. This simple definition of a corpus arose in the course of developing a continuing professional development workshop for working translators and language editors, in which practical tasks based on genuine translation problems are combined with practice-and-theory grounded perspectives. To date we have experimented or worked with specialist corpora created for various medical sub-specialisms, engineering, financial reports, rock mechanics, association bylaws and eighteenth-century medicine. Our approach to solving a translation problem, once the source has been understood, focuses on exploring the target language directly, after we have ensured that:

a) We search texts that are restricted to those that can be strictly matched in terms of genre with our source text (e.g., respiratory medicine as it appears in journals or other collections for peer readers, known in applied linguistics as a discourse community (Swales, 1990); UK financial reports written for investors and other real users of the information).
b) We examine co-occurring text (co-text) in the specific knowledge area of the source text (e.g., genes or proteins in the context of small

cell lung cancer, or derivatives accounting in the context of financial reports).

With these givens, we can confidently explore and examine possible alternative solutions to our editing or translation problem. All we need is a suitable tool that will enable us to rapidly conduct linguistically relevant and creative searches.

Our target readers for this paper—the translators or revisers for whom a time investment in the corpus-driven approach is worthwhile—will be practitioners such as:

- a novice translator who has decided to specialise in a particular field;
- a more experienced translator who wants to shift from a generalist to a specialist market and who wishes to give consistently high quality output while taking a proactive stance to career building;
- a translator who has a steady, valued client in a specialist field; or
- translators working in a team that needs to converge in terms of domain-specific language choices.

So that readers can see what kind of questions a corpus can answer, we first briefly describe two easy-to-learn, intuitive tools for analysing a corpus. We then discuss the basic steps involved in creating a suitable corpus (focusing on issues of text selection, collection and storage). In resolving the issue of sampling adequacy (corpus size) in a practical way, we propose combining a stable, cleaned substrate corpus in a knowledge field coupled with a more rapidly compiled ephemeral or ad hoc corpus to add greater topic specificity. We close with a brief discussion of whether the web can be considered a corpus and a reminder of why this approach must be distinguished from open-ended web searching.

2. Two corpus analysis tools: a concordancer and an indexer

Once good models for the target text have been collected and saved in a directory (i.e., a folder in a Windows environment), they can be analysed using a concordancer,[2] which works best when the corpus is composed of plain text (*.txt) files. If the corpus is composed of other file types (PDF, Word, HTML, etc), these can either be converted to plain text or analysed directly using an indexer.[3] The main practical difference between the two tools is that the former requires a time investment in pre-processing and cleaning up the files (an effort which ultimately pays off in more refined search outputs), while the latter requires only that the user store the model texts (Word documents, HTML files, PDFs) in a folder.

2.1. A concordancer

For translation purposes, the most intuitive, immediately useful feature of a concordancer is an output called a keyword-in-context (KWIC) display,

also called a concordance. Figure 1 illustrates such a display, which consists of a list of occurrences of the keyword (or phrase) and its co-text or span (i.e., around 10 or 15 words to the left and right of the keyword). The concordancer we use, called AntConc,[4] was developed for use by autonomous learners of English for specific purposes (Anthony, 2006). This tool, which has a highly intuitive interface, centres and highlights a keyword or words. It also has a function for right- and left-sorting concordances (Figure 2) to check for collocates and to test or confirm hypotheses in relation to how a word or phrase is used. Varantola (2003) discusses a wide range of problems student translators solved with another concordancer like this one. Wilkinson (2005a) shows further examples of concordancer outputs; they are particularly interesting because they are from a corpus of travel literature and so reveal that this approach is useful even for apparently simple translation specialisms. Although many might consider such an area not to be a specialism at all, Wilkinson reveals how the production of any text type that can be characterised, and for which a well-defined corpus can therefore be built, will benefit from this approach.

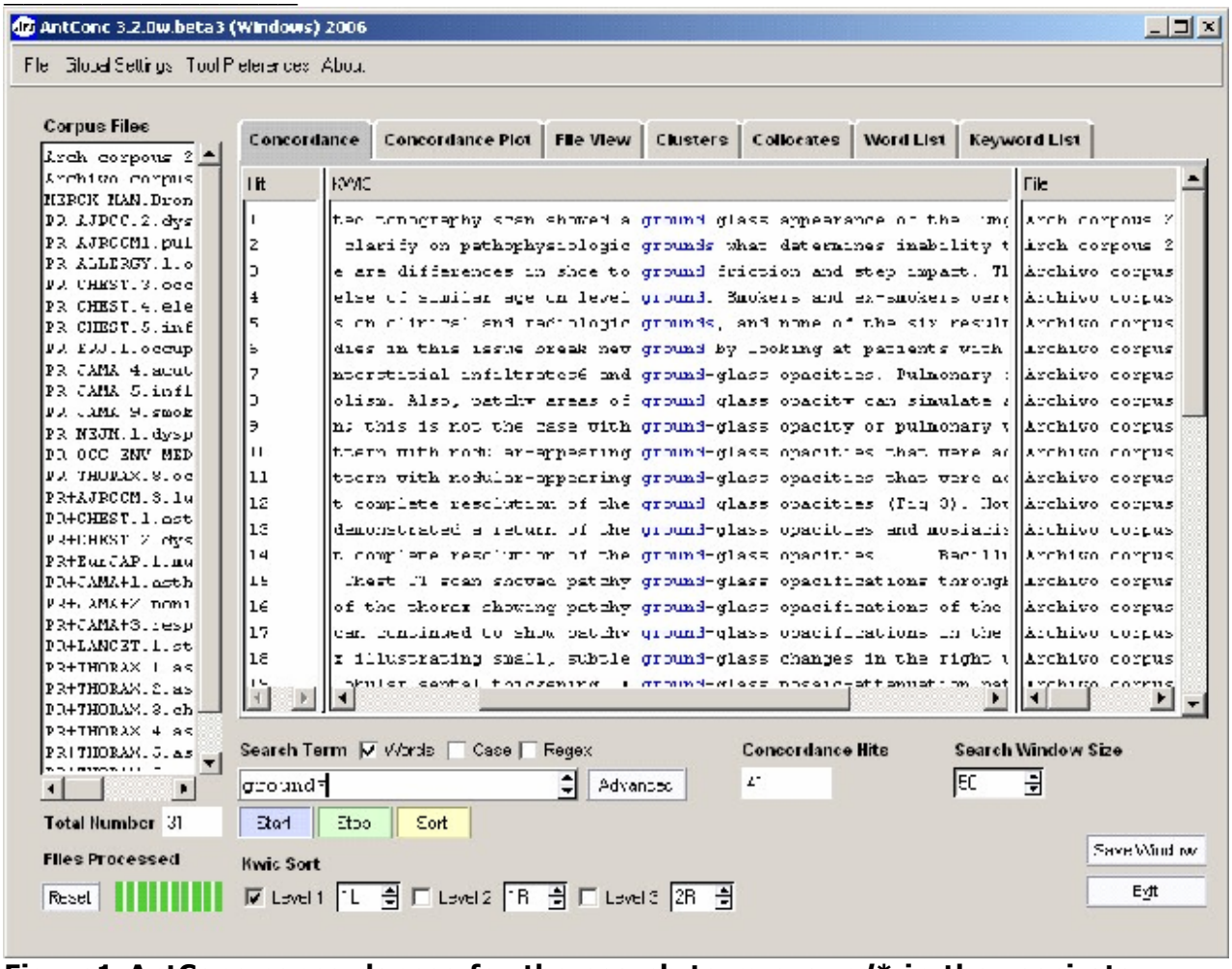


Figure1 AntConc concordances for the search term *ground in the respiratory medicine corpus. The asterisk can be used to reflect inflections or to replace full words occurring between, before or after other words.**

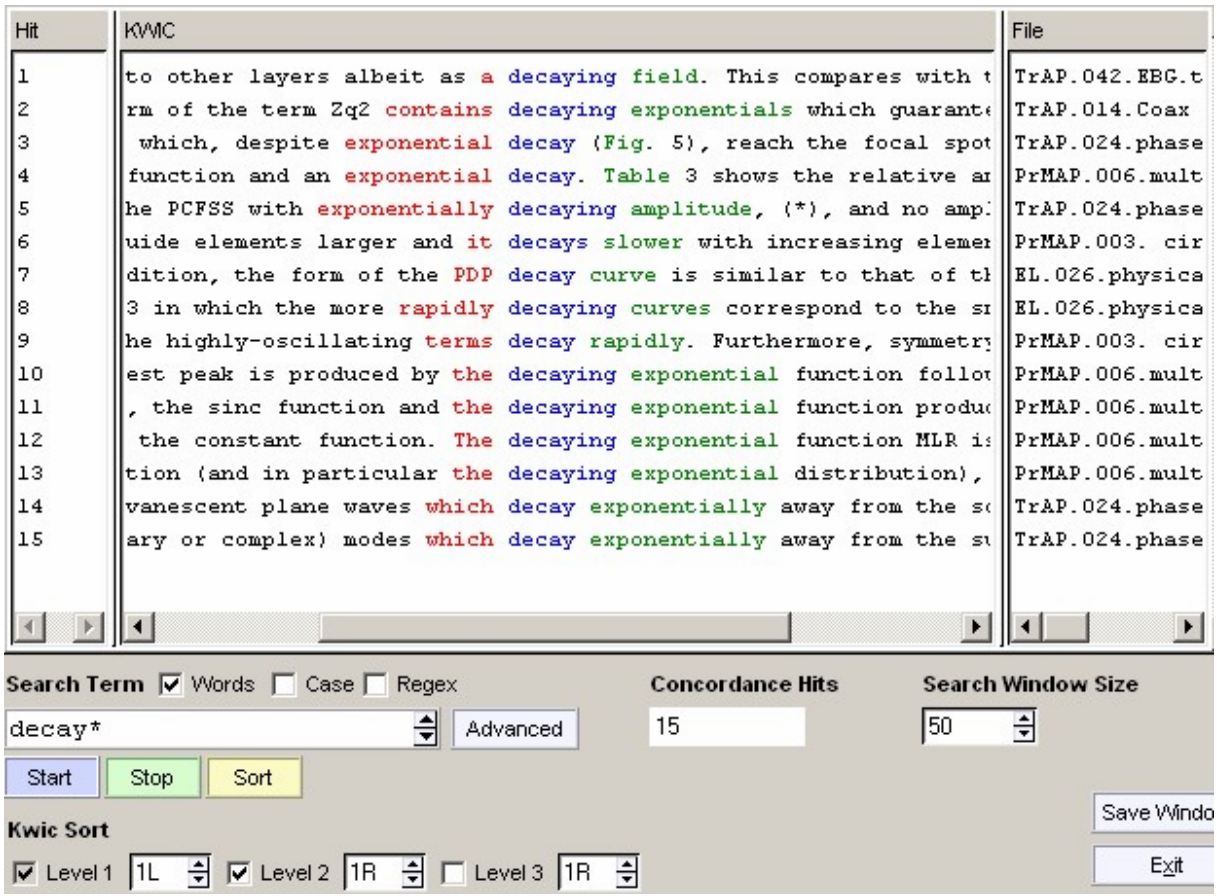


Figure 2. AntConc output for *decay in the signals and antennas corpus, with a 1-right sort (green)/1-left sort (red). The sort function helps identify patterns. The translator was asking two questions: a) What words might express the notion of a reduction in an exponential function (forms of *reduce* and *decay*)? and b) What adverbs might be appropriately used as intensifiers?**

A concordancer also allows a greater amount of context to be quickly examined (the file view function in AntConc). A word list function rapidly provides information on corpus size (number of tokens and types[5]) and gives a perspective on the salient features or 'aboutness' of a corpus by ranking words by frequency. This feature can be used to guide a translation team; it can also be used to extract keywords from a source text and guide a search for texts for a corpus. Other concordancer functions—of use mainly to researchers who analyse large corpora or educators—are the 'collocate' and 'cluster' functions (which tell us about 'the company a word tends to keep') and a keyword identifier that works by comparing one corpus to another.

2.2. An indexer

Although corpus analysis has lately become synonymous with concordancing, a corpus is not necessarily defined by the storage of texts in any particular form; any collection of model texts defined by appropriate criteria is a corpus. And any such collection, provided it is in digital form, can be mined with an indexer. We use a desktop search

application called Archivarius[6] that most users will find easy to use because of its Google-like output (Figure 3). Its text analysis functions are more limited than those of a concordancer as far as analysing syntactical patterns is concerned, but that may matter little to the translator who wants to mine downloaded texts immediately without having to pre-process them in any way.

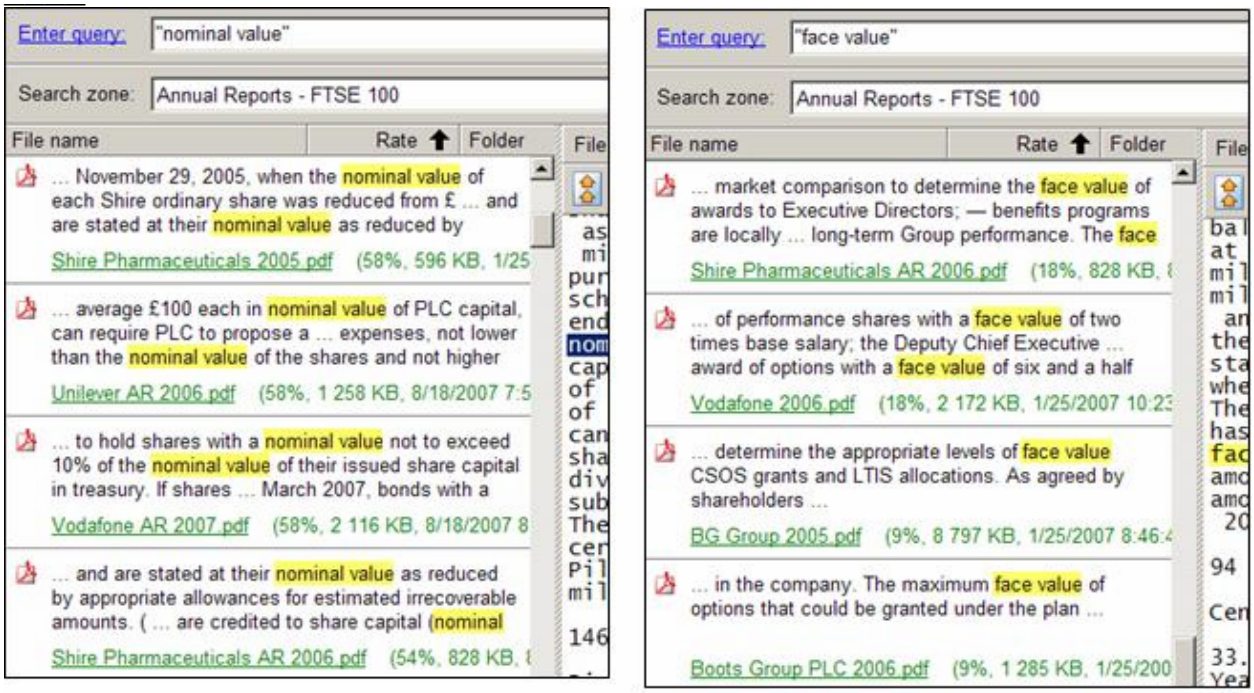


Figure 3. Two superimposed sample Archivarius outputs for the search terms *nominal value* and *face value* in the annual accounts corpus. When the user selects any of the Google-like hits on the left hand side of the Archivarius screen, the relevant section of a document is automatically displayed on the right side of the screen as simple text.

Since an indexer facilitates early adoption of a corpus-guided approach to translation, it is of immediate benefit to those who may still be uncertain as to the area in which they will specialise or who are still exploring whether conversion to text files (required for using a concordancer) is worth the effort. Many working translators already have a collection of model PDF or HTML texts associated with past work for regular clients, so using an indexer to mine such material merely requires grouping the texts in a single folder for indexing purposes.

3. Corpus creation

From the working translator-reviser-editor's point of view, the practical steps in a corpus-guided approach are to 1) accurately identify the type of text needed and find a source, 2) collect a sufficient number of the right text type, 3) store them in an appropriate form for analysis, and 4) analyse the language components. The previous section dealt with step

4—describing tools with which corpora can be analysed. We will now look at steps 1 to 3.

3.1. Identifying model texts

Corpus design—identifying the right content—is the key to confident decision making later. Failure to define the desired text model accurately can ultimately lead to translations that sound 'off' to the target reader. A client does not want a research article which sounds like a patient education pamphlet. Nor does the client want an annual report that sounds like financial journalism. Off-register error can also occur in the opposite direction: a client wanting the translation of a patient education pamphlet will not want to see *mucosa* used to refer to the lining of the nose and trachea—and that is what is likely to happen if a research article translator switches to patient material without consciously choosing a different model. We discourage broad sampling of the Internet by topic keywords alone if working to a high standard within a specialism. We recommend attention to genre and a discourse community's reading preferences, given that our goal is not primarily to find 'a good explanation of subject matter' (Lopez-Rodriguez and Tercedor-Sánchez, 2008), a purpose for which other research strategies can be equally effective. Rather, we wish to open a window that allows us to observe a community's language use. To define a corpus useful for that purpose, we ask these questions: how does the end user define the characteristics that set apart the texts we want to emulate from other texts on the same topic? Where are such texts to be found?

A translator familiar with the client's discourse community may well be able to answer these questions unaided. We created a respiratory medicine corpus of half a million words that we eventually came to refer to as a 'foundation' or 'substrate' corpus (explained in more detail below) to guide a team's translation of research articles, review articles and case reports. This corpus was created on the basis of our own direct experience of the journals most highly valued by academics in this field and in medicine overall. We knew how to identify peer-reviewed journals (see Gile and Hansen, 2004, for a discussion of academic peer review from the translator's point of view). Furthermore, within the peer-reviewed journal domain, we knew how to identify quality journals based on impact factor, indexing, editorial board prestige and other criteria. We recognised differences between these journals and industry-sponsored pseudo-journals or website look-alike content. Knowledge of which article types are typical in medicine also came from our own translation and reading experience. However, if we lacked familiarity with a discourse community, we would be guided by the reference sections of the articles to be translated. The fact that references are provided in such texts is, in fact, a distinct advantage for academic translators. In the field of finance, on the other hand, where references are not a systematic feature of texts, we were also able to quickly compile a million-word corpus to guide

financial report translations based on a reliable list of the UK's biggest publicly quoted firms (the FTSE 100).

A translator who cannot characterise the scope of the text types he/she requires will need an informant—an expert who can confirm that the translator's impressions about relevant corpus content are accurate or complete enough and provide guidance on what a discourse community values. In compiling a quarter-million-word corpus for antennas and signals engineering, we first compiled a list of relevant candidate peer-reviewed journals and then asked a senior researcher to validate our choices and to inform us as to article types in this field. A rock mechanics corpus was similarly created on the basis of a client's input. Such consultants can be used either to establish corpora, as in our last two examples, or to verify that corpus-based observations seem accurate to real members of the community (e.g., Anthony, 1999).

We also grappled with the question of whether or not to choose texts written by non-native speakers of English. In finance, we chose publications by major companies that were likely to have been professionally produced by teams of native speakers and communication companies. In medicine and rock mechanics, our corpora lean towards native speakers' texts, but must necessarily contain prose by non-native English speakers in fields where such scientists lead a branch of research. Although speakers of English as an additional language (whose articles are labelled E2 in our corpus logs) may provide very adequate help with specific terminology, not all parts of their texts may offer appropriate models.

Finally, a word must be said about dating texts. Corpora need to be updated because language changes over time. The more modern term *CT scan*, for example, would not have appeared as often as the now outmoded term *CAT scan* in a medical corpus closed in the mid-1990s. Furthermore, some jobs may require diachronic comparisons, making it important to log the dates of items in a corpus. Recently, it was necessary to carefully compare our eighteenth-century English corpus with texts from the middle of the nineteenth century. The source text from the Spanish Enlightenment discussed public and workplace health a good half century before the English public health movement gained force in the 1830s and 1840s with the work of Edwin Chadwick. Many English expressions now associated with that pre-germ–theory era come from Chadwick's period and tend to suggest the evident smells of vapours. The Spanish writer used expressions that suggested the essential changes of those vapours (described with forms of *corrupción*) rather than their manifestation (smells). Had the later expressions been adopted—particularly the term *putrid*—the translation would have made the Enlightenment author seem to be speaking off-century. This potential error could be avoided through diachronic analysis of properly dated corpus material.

3.2. Collecting texts

Forty years ago a linguist's corpus might have been a collection of facsimiles or a stack of books set aside in a university library carrel. Twenty-five years ago a corpus might have been a set of photocopies. And 10 or 15 years ago it might have been a batch of photocopies or originals to be scanned and digitised. Today, however, the great availability of texts in digital form undoubtedly facilitates the corpus-guided approach to translation and editing. Three digital collection issues need to be taken into account, however, if corpus building is to be useful: a) access to free, readily available texts, b) sampling adequacy, and c) fair use.

3.2.1. Access to material

Translators in some fields are more favoured than others when it comes to free access to texts that are used by insiders in a discourse community. Academic medicine is particularly well served. Many high-quality sub-specialist journals provide open access after six months or one year, and the main general medical journals have similar policies. Even journals that limit access to subscribers allow texts to be plucked as free-access editor's choice articles. Certain medical publishers, such as Biomed Central, are entirely open access to readers. Other academic fields might be slightly more difficult to sample, but hardly impossible. Many journals, for instance, will give free access to one sample issue. Harvesting several journals in this way, plus editor's choice offerings, should yield a small starter corpus. For certain fields—our engineering corpus was one example—a university access key or a visit to a university library will be necessary.

Harvesting appropriate texts in non-academic fields—or defining hidden corners in those fields—will require more creativity. Our corpus of FTSE 100 annual reports, for instance, was obtained by downloading free reports from company websites. User manuals for medical equipment, in contrast, were found to be largely inaccessible, although manufacturers have given us files willingly when we explained how we planned to use them. Certain text types—belonging to what Swales (1990) has called 'occluded' genres (never published and only seen by insiders)—are almost impossible to sample quickly and so best left to researchers in applied linguistics: reviewer and editor reports are an example, and researcher point-by-point responses to these reports are another. Access to these and many legal documents may require an insider's assistance, and even then there may be questions of confidentiality to be resolved; safeguarding anonymity may mean that the effort is not worth it except for translators who are fully dedicated to that sub-specialism.

3.2.2. Sampling adequacy

How large should a corpus be? This is an issue that speaks directly to those of us who must trade off an investment in time against longer-term benefits. A major reason translators or editors might choose to be guided by a corpus is because they wish, in the vaguest possible terms, to emulate the language of the domain; over the longer term, however, a wise translator begins to realise that using a corpus helps correct idiolect and reduces the possibility of over-generalisation from limited personal experience with language varieties. A corpus pulls together a broader set of models, reducing the temptation to rely on selective recycling of salient phrases that are sometimes too long and may leave an author open to accusations of plagiarism or cut-paste writing (Kerans, 2006). A corpus that is too small can lead, like personal experience, to skewed language choices.

We have been unable to locate a frank discussion of corpus size applicable to our working context, and are therefore still attempting to devise and validate a way to plan size in advance. However, after years of working with different-sized corpora, we have come to the conclusion that although size may affect the number and type of questions we can answer when mining a corpus, over-worrying about size may prevent wordface workers (translators, editors, language instructors, etc) from getting started at all. We must therefore say something about it.

Early on in our practice we observed that while a corpus as small as 40,000 words proved adequate to temporarily guide instructors entering in a new field of specialised language teaching, it was much too small for translation purposes. Yet the million-word corpus linguists often assume to be a minimum goal may be too time consuming to create (particularly if it is to be cleaned of artefacts and logged, as we recommend in section 3.3 below). By way of example, we mention that harvesting, converting and superficially cleaning a million-word eighteenth century prose corpus required a full day's work by an experienced corpus builder. The reason this was deemed worthwhile was that it would guide the translation of a book of 35,000 words into a form of English spoken by no living persons; the project, furthermore, required consensus between the translator and an expert editor (Kerans & Stone, 2008).

We advise novice corpus builders to quickly compile about a quarter of a million words and observe what kind of responses they get for questions posed. We found that this was the point at which our respiratory medicine corpus, for example, began to provide sufficiently useful answers to guide a team of translators converging toward shared practice. This corpus became even more useful when its size was doubled to half a million. At this point, however, it became clear that we would need to solve the problem of insufficiently broad scope. The logical solution, choosing highly topic-specific texts for addition to the core corpus, was an approach that

would require time for detailed analysis. It was then that we introduced a more practical concept, that of using the half-million word corpus as what we have come to call a 'substrate' corpus because it provides a firm base for more ephemeral corpora. A substrate corpus contains carefully chosen, logged texts that have been cleaned of non-linguistic artefacts to a high standard so that it can reliably provide both frequency counts and information about the collocation patterns that give a specialist language its underlying form.

With this concept in place, once we accepted that there was a practical size limitation with regard to building such a clean substrate corpus of good models, we were open to the notion of adding what Tribble (1997) described as 'quick-and-dirty' (Q+D) corpora, meaning 'small, informally produced corpora'. Although Tribble was encouraging specialist language instructors to study such small corpora rather than rely on instinct, the phrase has come to be used to characterise any corpus rapidly harvested from the Internet, but not cleaned or logged for systematic safekeeping and building. Ideally, the texts for a Q+D corpus will offer models for usage that are of similar quality to those of a substrate corpus in terms of genre appropriateness. The uncleaned corpus will simply give a more haphazard-looking output, or may include duplications, making some concordances more difficult to interpret.

Our Q+D corpora mainly serve to enhance topic range quickly, solving the small-corpus problem of inadequate sampling. In other words, we have invested the necessary time and effort in building very clean substrate corpora for respiratory medicine and other fields, but, on a job-by-job basis and only if needed, we supplement them with terminologically rich Q+D corpora created for specific topics. It is possible to do the harvesting in a highly automated manner. One translator on our medical team has added as many as another million words within minutes using an online corpus creation tool (WebBootCat[7]). Other team members have manually gathered as few as an additional 40,000 words of highly specific prose or up to 150,000 new words on an emerging topic or a new research design. Such enhancement is necessary only when a particular job requires terminology within a narrow topic range. Many translation commissions are adequately guided by a substrate corpus alone, however, so we are free to load an additional Q+D corpus or not, as we see fit.

3.2.3. Fair use

The question of fair use refers to the legality of collecting, storing and using texts without first obtaining permission from copyright holders, an issue discussed in detail from a practical corpus-for-translation perspective by Wilkinson (2006b). The main problems arise not with using a corpus for personal reference purposes but with two related circumstances in working translators' and academics' lives: a) reproducing

extracts in research articles (e.g., as KWIC displays like those in this article), and b) sharing corpora with colleagues.

In regard to the first of these issues, according to Davies (2002, as cited in Wilkinson, 2006b), the copyright law that matters is that of the country from which the corpus is distributed and not the country in which the texts were created or in which the corpus user accesses the material. We are uncertain of Spanish law in regard to the use and reproduction of corpora. However, our position is that when we reproduce figures such as those in this paper, we are not citing the ideas in the specific texts. Rather we are displaying language patterns that are not specific to the usage of particular authors; as the concordance reveals, they are more generally applicable patterns. Hence, citation of the original authors' work is irrelevant, though technically possible in our logging system.

Wilkinson (2006b) also states that the fair use issue is even 'murkier' with regard to sharing corpora. At present, we share corpora with a clear conscience; when making a corpus freely available to translation team members or colleagues through a non-profit professional association's workshops,[8] we do so in good faith and feel no harm is done. The receiver's use is personal, and our practice is analogous to a university professor sharing medical articles with students. Note, moreover, that for many fields for which a corpus might be created, the issue is moot: the annual reports in our financial corpus are all freely and widely distributed and carry no copyright statement at all.

To sum up, we feel that the technical capability for creating and analysing useful corpora is far in advance of the law's awareness of the practice. In the absence of clear instructions, our need to know about these tools and put them to use in benefit of our clients and their readers takes precedence. By way of contrast, however, we mention the more careful approach of the Professional English Research Consortium (PERC),[9] which is compiling a 100-million word corpus representative of several knowledge and practice fields. The PERC anticipates that the corpus—in fact several sub-corpora—will eventually be used by language researchers under license; they are therefore carefully soliciting and obtaining permission from copyright holders.

3.3. Preparing and storing texts as a corpus

In one sense, corpus storage merely means placing a collection of texts in a directory or folder. When storing texts for processing in a concordancer, original format files (PDFs, HTML documents, etc) can be stored alongside text files conveniently in the same folder and with the same names, as AntConc will only load the files with the *.txt extension.

Certain decisions about file labelling and storage can save time over the long run and facilitate problem solving, however. And most importantly (as seen in Section 2 above on corpus analysis), how we decide to save files affects which tool we can use. This section will cover: a) why and how we log and label files systematically; b) the merits of various ways of creating text (*.txt) files; and finally, c) how to clean up a text file to varying degrees—and why it is worthwhile to do so—versus when to simply work with Q+D corpora.

3.3.1. File names and logs

File names should give information about the provenance of a hit at a glance, so that the editor or translator who knows the content of a corpus can factor in that information when judging suitability. Above, Figure 2 shows AntConc file names in the right-hand column; Figure 3 shows Archivarius file names—referring to companies—in green under each hit listed on the left of the output screen. Codes known to the translator or editor who uses the corpus give information on—for example—which academic journal published a text (e.g., *PrMAP* refers to *Proceedings: Microwaves, Antennas and Propagation*) followed by the main topic. It is useful to look at and learn from poor file labelling too: note that Figure 1 shows hits only for two files whose names provide very little information. The first file is a set of website texts from hospitals and university programmes containing instructions on how to perform diagnostic and surgical procedures, whereas the second file—referred to with the tag PR (indicating that it is peer-reviewed)—is a set of texts from formal medical journals. Although both contain texts written by professionals for other professionals, they represent different genre families. These file names—dating back to our early corpus creation period—provide too little information about text provenance and topic. It is still possible to go to the File View option to check those files, but a working translator wants to make informed decisions quickly. Therefore, with later additions to the corpus, our labelling became more enlightened. Now, the informed user of our corpus assessing the merits of an item in a KWIC display can immediately know topic, whether a text is British or American, or whether an author's first language is not English (E2). Suitably informative file names are also useful in another way: they permit an editor or translator to load files selectively from overlapping sub-corpora and so make corpus coverage wider or narrower.

Logging corpus content (Figure 4 shows a simple example in Word, although we are now beginning to use databases) may seem unnecessary, but we have found that doing so avoids duplicated effort if a corpus is shared by a team or if a lone translator keeps and updates several sub-corpora in a single specialism. Valuable time is obviously wasted if the same file is prepared by more than one person or more than once.

	File name (Txt + Intact versions)	Type / genre	Citation information	Words	Key words
1.	MERCK MAN.Bronchiectasis	BOOK online	Merck Manual, section 6, chapter 70	3727	
2.	PR AJRCCM.1.pulmonary rehab	IMRD	RIES, Andrew L., Robert M. Kaplan, Roseann Myers and Lela M. Prewitt. **Maintenance after Pulmonary Rehabilitation in Chronic Lung Disease: A Randomized Trial** *American Journal of Respiratory and Critical Care Medicine* Vol 167. pp. 880-888, (2003)	4634	
3.	PR AJRCCM.2.dyspnea	IMRD	MOY, MARILYN L., Mary L. Lantin, Andrew Harver, And Richard M. Schwartzstein **Language of Dyspnea in Assessment of Patients with Acute Asthma Treated with Nebulized Albuterol** Am. J. Respir. Crit. Care Med., Volume 158, Number 3, September 1998, 749-753	4124	physiotherapy, 6-min walk test, treadmill, exercise, quality of life
4.	PR AJRCCM.3.lung function (Invaluable resource on variables and measurement procedures.)	CONSENSUS STATMENT	**ATS/ERS Statement on Respiratory Muscle Testing** Am J Respir Crit Care Med Vol 166. pp 518–624, 2002	75,347	
5.	PR ALLERGY.1.occup asthma	Editorial	J.-L. Malo **Occupational rhinitis and asthma due to metal salts** Allergy Volume 60 Page 138 February 2005 Issue 2	843	rhinitis, metal salts
6.	PR CHEST.1.asthma	CASE	WECHSLER, Michael E., MD; David Finn, MD; Dineli		neurological

Figure 4. A log as a Word table. Databases or spreadsheets can also be used. This log contains a short but immediately informative file name (used for both the original-format file and the text file). It also describes the genre (article type), and provides bibliographic information to ensure the entry will not be duplicated, a word count, and additional keywords.

3.3.2. Conversion to plain text

Files are saved both in their original format (usually PDF or HTML) and with the *.txt file extension, and under the same names. The original format is a readable document that is useful for examining tables and figures or for learning about content. This version is also useful in order to be able to correct any errors that occur during conversion and cleaning.

The text (*.txt) format that is a standard requirement for conventional concordancers can be obtained in a variety of ways. Some documents can be directly downloaded from the web as off-copyright e-texts (e.g., from the Project Gutenberg or similar repositories). For some specialisms, adding e-text to your search string can locate useful book additions in a very clean form.

Many specialisms are best served by PDF or HTML collections, however. A feasible procedure is to convert texts using the browser's 'save as' option (choosing 'text file' from the sub-menu) or using Acrobat Reader's 'save as text' option. Cleaning such files can be time-consuming, however. Coding artefacts must be removed and the converted text proofread to rectify jumbled lines or paragraphs. If you must use this option, we recommend converting from the HTML version as the cleaning and checking process is usually easier. A much better conversion can be obtained by using a commercial PDF file converter—a small but worthwhile investment for a corpus-guided translator or editor.[10] The resulting text files are almost ready to use, and how much more work you do depends on the level of cleaning you need.

3.3.3. Cleaning files—is it necessary?

In our experience, minimal clean-up of a well-converted plain text file (with content in the correct order) is necessary, at least for a reliable substrate corpus that shows patterns faithfully. Cleaning enriches outputs because it ensures that a search will include all instances of a word or phrase and will not exclude occurrences because of a punctuation, spacing or coding anomaly. Here are the basic steps to follow:

- **Remove reference lists (if present).** Although you have chosen a text as a model to emulate, you have not chosen each of the references used by an author. Hits from titles in the references section (chosen on the basis of non-linguistic criteria) can distort frequency counts and introduce non-preferred usage.
- **Remove non-linguistic content.** This step may be unnecessary if a good PDF converter has been used. If HTML documents have been converted directly from the browser, the beginning and end will have large blocks of coding. In both cases, leave only sufficient labelling at the beginning of a file to allow easy identification of the source. Remove coding for most tables and figures but leave legends and titles, as these often have useful language information.
- **Remove extra spaces.** Failure to remove extra spaces can skew frequency counts. A search for a two-word string like *pathology report*, for example, will not include clusters that contain more than one space between the two words. Note also that apostrophes are also sometimes followed by unwanted spaces after conversion.

More exhaustive cleaning involves the following additional steps:

- **Correct words that appear with anomalous characters or symbols** (often denoted by a question mark).
- **Correct problems related to hyphenation in the original text.** Sometimes words at the ends and beginnings of lines in the PDF are joined together or broken up. Correct these and also remove any discretionary hyphens (marked by the symbol ¬) that may be present.

Opening a text file in Word and using the search and replace options can make cleaning easier. Switching on the spell/grammar check function also helps locate anomalous artefacts. Before saving the file as a text document, check that it includes, at the head of the file, the bibliographic information that identifies it.

Finally, remember that for logical cost-benefit reasons, translators need to be sensible about the levels of cleaning thoroughness to apply.

4. Discussion and conclusions

Using corpora to guide translation or editing work is a way to compensate for any or all of the following: a) uneven field knowledge; b) non-contact with language genres and registers outside our normal range of use; and c) source language interference from lack of contact with our native language. In general terms, using corpora can help us mature as specialist language users.

We described two tools that can be used for analysing corpora. Although the search possibilities for studying collocations with the concordancer (AntConc) are more sophisticated, the indexer (Archivarius) has the advantage of enabling searches of a variety of text formats. The indexer, therefore, allows a corpus-guided approach to be applied when, for practical reasons, we need immediate corpus research capability and may already have model texts to hand. Over the long term, however, the serious specialist will require the sophistication of a concordancer to be able to address trickier language issues.

Irrespective of which tool we prefer to use at any given time, however, we cannot emphasise enough that building a successful specialist translator career on the basis of corpus-guided translation or editing largely relies on the quality of the substrate corpus. This is not to say that uncleaned, topic-oriented corpora do not have their uses. We previously referred to a hierarchy that can range from time-consuming manual corpus creation to instant and automated corpus building with a web-based tool fed with keywords. The different approaches are complementary and can be combined, and in some cases, a web-harvested corpus alone may be adequate for certain subject areas or tasks (as found when we created a bylaws corpus to guide the translation of an association's charter or when a colleague translated an oceanography website). A rough-and-ready corpus must be mined with care, however, as it has sampled the wider web's many genres indiscriminately. To quote John Sinclair (2004):

> The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present it is quite mysterious (…) and it is not at all clear what population is being sampled. Nevertheless, the WWW is a remarkable new resource for any worker in language.

We agree that the availability of a vast range and quantity of digital texts that can be rapidly harvested off the web is a key factor underpinning the current practicality of the corpus approach. Success, however, requires using appropriate models to minimise errors of style, register and terminology. There is no substitute for applying well-considered human criteria to the creation of a reliable, well-characterised specialist corpus in which we have confidence when making decisions. The serious specialist

who is ready to benefit from a judiciously compiled and very clean substrate corpus will have to deal with the three issues we have raised: ease of access, which varies according to knowledge field; corpus size, for which guidelines are still needed if time investment is to be kept under control; and fair use, for when a corpus is likely to be shared. The obstacles thrown up are surmountable, however, and the short- and long-term rewards considerable.

References

- **Anthony, Laurence** (1999). Writing research article introductions in software engineering: how accurate is a standard model? *IEEE Transactions on Professional Communication.* 42(1); 38-46.

- **Anthony, Laurence** (2006). Developing a freeware, multiplatform corpus analysis
- toolkit for the technical writing classroom. *IEEE Transactions on Professional Communication.* 49(3); 275-286.

- **Gile, Daniel** and Gyde Hansen (2004). The editorial process through the looking glass. In: Gyde, Hansen, Kirsten Malkmjær and Daniel Gile (eds). *Claims, Changes and Challenges in Translation Studies.* Amsterdam/Philadelphia: John Benjamins;297-306.

- **Hunston, Susan** (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

- **Kerans, Mary Ellen** (2006). Avoiding innocent plagiarism—the plagiarism of *innocència* by authors and their language consultants. Online at http://www.metmeetings.org/old_content/pagines/METM2006/Presentations/Presentation%20-%20Mary%20Ellen%20Kerans.htm (consulted 24.02.2008)

- **Kerans, Mary Ellen** and John Stone (2008). Notes on the translation and editing of Masdevall's *Account of the Epidemics* and his *Opinion* on health in the textile industry [Translators' notes]. Masdevall Joseph. *Relación de las epidemias de calenturas pútridas y malignas. Account of the epidemics of putrid, malignant fevers afflicting the principality of Catalonia in recent years; chiefly concerning their discovery in the year 1783 last in the city of Lerida, the plain of Urgel and many other administrative districts and divisions; including the successful, quick and certain method for curing such diseases*. Barcelona: Ars XXI, 55-61.

- **López-Rodríguez, Clara Inés** and María Isabel Tercedor-Sánchez (2008). Corpora and students' autonomy in scientific and technical translation training. *Journal of Specialised Translation*, Issue 9, 2-19. Online at: http://www.jostrans.org/issue09/art_lopez_tercedor.php (consulted 29.02. 2008).

- **Moreno, Ana I.** (1997) Genre constraints across languages: causal metatext in Spanish and English RAs. *English for Specific Purposes.* 16(3), 161-79.

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

- **Sinclair, John** (2004). Corpus and text — basic principles. Martin Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow Books: 1-

16. Online at http://ahds.ac.uk/guides/linguistic-corpora/chapter1.htm (consulted 27.02.2008)

- **Swales John** (1990). *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

- **Tribble, Chris** (1997). Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. Melia James and Barbara Lewandowska-Tomaszczyk (eds) (1997) PALC 97 Proceedings, Lodz University Press: Lodz, 106-117.

- **Varantola, Krista** (2003). Translators and disposable corpora. In: Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.) *Corpora in Translator Education*. Manchester: St Jerome.

- **Wilkinson, Michael** (2005). Using a specialized corpus to improve translation quality. *Accurapid* Vol 9 (3). Online at: http://www.accurapid.com/journal/33corpus.htm (consulted 27.02.2008)

- **Wilkinson, Michael** (2006). Legal aspects of compiling corpora to be used as translation resources: questions of copyright. *Accurapid* Vol 10 (2). Online at: http://www.accurapid.com/journal/36corpus.htm (consulted 27.02.2008

- **Williams, Ian** (2005). Thematic items referring to research and researchers in the discussion section of Spanish biomedical articles and English-Spanish translations. *Babel* 51:2, 124–160.

- **Williams, Ian** (2006) Towards a target-oriented model for quantitative contrastive analysis in translation studies: an exploratory study of theme-rheme structure in Spanish-English biomedical research articles. *Language in Contrast.* 6(1), 1-45.

- **Zweigenbaum Pierre** and Natalia Grabar (2003). Corpus-based associations provide additional morphological variants to medical terminologies. *American Medical Informatics Association Annual Symposium Proceedings 2003*; 768–772.

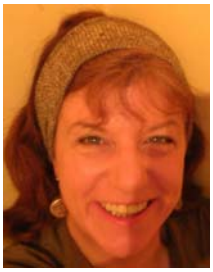
Acknowledgement

This article developed out of the workshop entitled 'Corpus-Guided Editing and Translation of Specialist Texts', first piloted in Barcelona in July 2006, offered again in Barcelona in May 2007, and in Madrid in October 2007. It will run again in Split, Croatia, on 10 September, 2008. This workshop is part of MET's expanding continuous professional development programme. For more details, visit http://www.metmeetings.org/.

Biographies

Ailish Maher, a freelance translator and occasional author's editor, holds the Institute of Linguist's Diploma in Translation and an MA in Translation Studies. Her thesis was on the subject of acquiring specialist macroeconomics expertise using a purpose-built corpus. She is Training Chair of Mediterranean Editors and Translators (MET).

Email: gaebolga@gmail.com



Stephen Waller, a freelance translator specialising in business and finance, has a degree in German and French. His interest in corpus-guided translation comes from the experience, common to many translators, of having to write convincingly about a wide range of specialist subjects.

Email: swaller@mailforce.net

Mary Ellen Kerans, a biomedical translator and author's editor, received her MA in TESOL. Her career in special-purposes English instruction fostered her interest in applications of corpus analysis.

Email: METworks@gmail.com

[1] See Tim Johns' Kibbitzer 6: http://www.eisu2.bham.ac.uk/johnstf/revis006.htm.
[2] A concordancer works by aligning keywords—its most basic function—so that the other words occurring in the vicinity can be identified, patterns discerned and the meaning of frequencies assessed.
[3] An indexer works like Google or any search engine. A desktop indexer, however, will invite the user to establish collections of texts within folders, so it is particularly appropriate to a corpus-guided approach to translation.
[4] AntConc is freeware, available from http://www.antlab.sci.waseda.ac.jp/.

[5] A text with 100 words is said to have 100 tokens. However, because some words will be repeated, there may be only (say) 40 different word types in this text.
[6] Archivarius (not free but very reasonably priced) is a desktop search tool that we find particularly well suited to our purposes: http://www.likasoft.com/document-search/.
[7] WebBootCat is accessed via http://sketchengine.co.uk/, a site which incorporates a simple online concordancer. An annual subscription is required to use the corpus builder and concordancer. A 30-day free trial is available.
[8] Mediterranean Editors and Translators: http://www.metmeetings.org/?section=workshops.
[9] Readers can learn more about PERC and the CPE at the group's website: http://www.perc21.org/.
[10] Two popular file converters are Iceni Gemini (http://www.iceni.com/gemini.htm) and Abbyy PDF Transformer Pro 2 (http://www.pdftransformer.com/).