Full citation: McEnery, T. and Brookes, G. (2022). 'Building a written corpus: what are the basics?'. In: A. O'Keeffe and M. J. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics* (2nd edition). London: Routledge, pp. 35–47.

4 Building a written corpus: what are the basics?

Tony McEnery and Gavin Brookes

1. Written corpora: An introduction

Despite the expansion in the range of modern-day corpora, the field of corpus linguistics continues to privilege the study of written language over other modes. This is because constructing corpora of written texts tends to be considerably easier, cheaper and quicker than constructing corpora of other modes such as speech. Indeed, the ready availability of machine-readable language in an increasing range of written genres means that it has never been easier to build corpora of written language. As well as outnumbering corpora representing other modes, written corpora also tend to be much larger than their counterparts. For example, both the original and updated versions of the British National Corpus (BNC) contain around 90% writing and 10% speech (Aston and Burnard 1998; Love *et al.* 2017; Hawtin 2018).

All this does not mean to say, however, that the task of building a written corpus is necessarily easy, or without complication. This chapter explores the main considerations involved in building a written corpus and addresses some of the challenges that the builders of such corpora typically face. The distinction between writing, speech and other modes is at best blurry. For the purposes of this chapter, we adopt a broad view of 'written' language, taking it to include texts containing language that has been handwritten or typed, including using a computer or other digital device, but also texts that represent speech but which have not been transcribed for the express purpose of linguistic research, such as those found in the Old Bailey corpus of trial discourse dating from 1674 to 1913 (Huber *et al.* 2016). For corpora that represent spoken language in some form, like the Old Bailey example, we will focus on the design and construction of corpora using the written forms of such texts, rather than the collection and transcription of original spoken discourse (see Chapter 3, this volume). While we pay special attention in this chapter to issues of particular relevance to the construction of written corpora, many of the points raised will be applicable to the construction of other types of corpora, too.

2. Design

Being as clear as possible on what we want to do with our corpus prior to designing it is important, as not only will this factor into our decision-making throughout the corpus-building process, but it might also be the case that an existing, publicly available corpus is sufficient for our requirements. Once we are clear about our purpose in using a corpus, the distinction between specialised and general corpora becomes relevant. Specialised corpora are designed to represent a particular genre or variety of language, usually within a specified context and/or time frame (e.g. Brookes and Baker's (2021) corpus of newspaper articles mentioning *obesity*). General corpora, meanwhile, represent language use on a broader scale, often whole languages (e.g. the BNC), and tend to be much larger than specialised corpora. If we are interested in studying the use of a particular linguistic item or feature on a broad scale, then it is likely that an existing general language corpus will meet our needs. However,

if we are interested in a particular variety or genre, or language around a particular topic or within a particular (recent) time period, then it is likely that we will have to build our own corpus.

Should we need to build our own written corpus, the decision regarding which texts to include in it can be quite straightforward and may be determined by our research question (e.g. with so-called 'closed' or 'finite' text types, such as the known literary works of a deceased author). However, in many cases it is not possible to include all relevant texts in a corpus. This is because corpus builders tend not to have access to, or even knowledge of, the full extent of the texts that could be deemed relevant to their research purposes, however well-defined these are. Most corpora therefore constitute *samples* of all possible candidate texts. Designing a corpus essentially involves deciding on which texts will be included in this sample and whether we will include these texts in their entirety or sample material from them. These considerations touch upon the related considerations of authenticity and representativeness.

Authenticity

The concept of *authenticity* in corpus linguistics research denotes a quality of language being 'naturally occurring' or as existing in the 'real life' (McEnery and Wilson 2001). Anything which involves the linguist beyond the minimum disruption required to acquire the data is reason for declaring our collection to be a special corpus, built with a looser interpretation of the criterion of authenticity. This opens up the possibility of special corpora composed of language that is semi-elicited by researchers through, for example, surveys or questionnaires, or that which is produced as part of a psycholinguistic experiment. In most cases, though, when designing a written corpus we usually want to include in it texts that can be judged as being as 'authentic' as possible, with minimal interference from the researcher. The task of capturing and including authentic language in a corpus tends to pose much less of an obstacle for the collection of written texts than it does for spoken texts (see Chapter 3, this volume).

Even so, researchers such as Mishan (2004: 219) question the capacity of corpora to represent 'authentic' language at all, claiming that the process of transposing texts into corpora 'forfeit[s] a crucial criterion for authenticity, namely context' (see also: Widdowson 2000). However, as technology develops some of this context-stripping is more a question of choice than necessity – to an extent, the use of annotation (discussed below) along with recent advances in the development of multimodal corpora and methods of multimodal corpus linguistic analysis represent means of recontextualizing the texts in a corpus, at least to an extent (see Chapters 3 and 7, this volume, for more on multimodal corpora).

Some issues surrounding authenticity are particularly pronounced in, and in some cases unique to, written corpora. One such issue relates to orthographic representation. Many types of written texts contain nonstandard spellings of words. Typical examples include: texts that have not been professionally edited, for example student essays; casually produced texts such as those found on social media sites; and other types of user-generated web content like blogs, wikis and personal websites. The issue of nonstandard representation is also relevant to the collection of historical texts which, however carefully produced, are likely to have been written using now-outdated spelling conventions. Texts such as these are not problematic in themselves but can present an issue for the purposes of producing automated frequency counts. We can address this issue, however, by standardising the spellings in our corpus – yet, unless this is done with due regard for authenticity, we

may compromise any claims regarding the authenticity of our corpus. We will return to consider this issue in Section 3.

Another issue pertaining to authenticity in the compilation of written corpora concerns the volume of online language that is produced by so-called social media 'bots'. Bots can automatically generate large volumes of social media content very quickly. Such content can also be difficult for the untrained eye to distinguish from content that is produced by humans. A 2018 study of 1.2 million tweets containing hyperlinks to external websites reported that 66% of the tweeted links in the sample were shared by suspected bots (Wojcik *et al.* 2018). Although bots are trained on natural language, the texts they produce clearly cannot be considered to constitute 'authentic' instances of language use, so will have to be accounted for and preferably removed if we are building a corpus of social media posts such as tweets, unless, of course, our goal is to generate a special corpus of bot tweets.

Representativeness: Balance and size

Once we have decided on the written language or variety we want to investigate, we must decide on which texts to include in our corpus. Unless we are building a very specialised corpus, we are unlikely to be able to include all texts belonging to the language or variety we are interested in. In this case, a general corpus will necessarily represent a sample of the totality of the language or variety that we are interested in studying. When deciding on which texts to include in our sample, we want to ensure that it results in a corpus that is as representative of our given language or variety of interest. A widely accepted definition of 'representativeness' in the context of corpus linguistics is provided by Biber (1993: 244), who describes it as 'the extent to which a sample includes the full range of variability in a population'. Here, 'population' refers to the 'notional space within which language is being sampled' (McEnery and Hardie 2012: 8).

For the design of some corpora, such as those comprising texts that are not so easily accessed, for example because they are written in a minority language (see McEnery and Ostler (2000) for a discussion) or to which we have been granted privileged access, the issue of representativeness is less pressing as we may want to employ a more opportunistic approach and collect as many texts as we can. However, for the majority of corpora representativeness is a central consideration and informs how many texts we want to include and in what proportions.

Biber (1993) suggests that in corpus design, 'variability can be considered from situational and linguistic perspectives, and both of these are important in determining representativeness. Thus, a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language' (*ibid*.). A corpus can therefore be considered representative if it matches the situational and linguistic variability of the population under study (see also Chapters 1, 5 and 6, this volume, on representativeness). These issues can be viewed as aligning to two further features of representativeness: *balance* and *size*.

Balance

The concept of *balance* refers to the internal consistency of a corpus in terms of the proportions that are contributed by each variable (situational and linguistic). To maximise representativeness, the balance of these variables should reflect both qualitatively and quantitatively the situational and linguistic variables that constitute our target population. The ideal of a balanced corpus that matches the make-up of our target population is, however, usually not met because our ability to judge the

representativeness of our corpus depends on a clear definition of the 'population' under study. Yet establishing such a definition can be extremely difficult, or even arbitrary, in practice. The main reason for this is that we are usually not aware of the full extent of the population that we are studying, let alone how variables distribute across it. For example, assembling a balanced corpus of written British English would require knowledge of all written British English throughout the course of history and how this writing distributes in terms of, amongst other things, the demographics of writers, the genres and purposes of different types of writing, and the times and places in which that writing took place. A related issue here is that, even if we have a rough idea of the breakdown of writing with respect to the above variables, we ideally also need to know how widely the texts have been received. Another challenge to balancing our corpus to match the distribution of the authentic language it is designed to represent is that there are particular contexts or types of language to which we, as (corpus) linguists, have limited or no access, e.g. texts produced in private domains.

Data availability can present a particular challenge for designing representative corpora of historical written texts. One issue is that the archives from which the texts for such corpora are sourced can be incomplete. Another more general issue relating to the compilation of historical corpora is that it can be difficult, and in many cases impossible, to identify the socio-demographic backgrounds of those who authored the texts we are sampling, particularly when those authors are anonymous. The anonymity of texts' authors is not a challenge that is particular to historical texts, though, as the anonymous nature of much online interaction, including social media texts, presents difficulties for demographically balancing any corpus containing such texts.

In view of these challenges, it is not surprising that most corpora (written or otherwise) are not balanced in a way that perfectly represents the distributions of their target populations. Balanced, representative corpora are best viewed as a theoretical ideal rather than as being necessarily achievable in practice, resulting in two broad approaches to text selection in corpus design, 'balanced' and monitor corpora. The 'balanced' corpus approach uses a sampling frame which ensures equal representation of a particular set of variables by dictating which types of language or texts should be included in a corpus and in what proportions. Such corpora are balanced in the sense that they provide an equal representation of a set of variables. They do not necessarily reflect the actual distribution of those variables in the target population that the corpus is designed to represent. The Brown family of corpora, based on the sampling frame established by Kučera and Francis (1967), are a good example of this. The sampling frame produces corpora which contain approximately one million words of written English prose in samples of approximately 2,000 words taken from 500 different text sources from four categories of writing (press, general prose, learned writing and fiction), which are further split into 15 subcategories or genres. Corpora such as LOB (Johansson *et al.* 1978), the Freiburg-Brown corpus of American English (Frown; Hundt *et al.* 1999) and the Freiburg-LOB Corpus of British English (FLOB; Hundt *et al.* 1998) and the AmE06 and BE06 corpora (Baker 2009) followed that basic sampling frame. The benefits of building a corpus according to a sampling frame is that it can then be compared more systematically against other corpora which were designed using that same sampling frame, and so contain identical proportions of texts belonging to each category and subcategory. Yet the balance they strive for is not necessarily representative of the proportions with which the variables in question occur in the real-world populations they nominally represent.

Monitor corpora continue to grow over time, but not according to the variables which produce 'balanced' corpora. Instead size and currency are given precedence over balance. Perhaps the best-known example of a monitor corpus is the Bank of English (see McEnery and Hardie 2012:80). The texts in this corpus began to be collected during the 1980s and now constitute part of a larger monitor

corpus, the 4.5 billion-word Collins Corpus. While it could be argued that monitor corpora produce an imbalanced view of the languages or situations under study (explored more in the next section), it might also be argued that any imbalance self-corrects over time due to the sheer size of the corpus and the fact that the skew in the texts included is not consistent. The concept of the 'Web as Corpus' (Gatto 2014) is, in many ways, comparable to the use of monitor corpora, as it takes as its starting point a massive collection of texts that continues to grow. In the case of Web as Corpus, this growth occurs independently of the researcher's efforts, as more and more user-generated content is added every second, potentially offering the largest and most up-to-date view on (most) languages that is possible at any given time.

The use of the Web as a Corpus, or at the very least the use of the Web for obtaining material to put into a corpus, seems to overcome many of the practical limitations that attend to the construction of corpora of written language derived from other domains. As Collins (2019: 32) points out, the Web is 'freely available, encompasses a breadth of texts from around the world and is of unimaginable size. Texts are already in a computer-readable format, so researchers do not need (on the whole) to undergo tasks of digitisation (as with written texts) or transcribing spoken data'. Indeed, because of the relative ease with which they can be collected and the minimal amount of processing they require for corpus analysis, corpora of Web texts are often very large in size. An example of this is the *TenTen* family of corpora (Jakubíček *et al.* 2013), which represent online texts written in a variety of languages. At the time of writing, the *Sketch Engine* tool (Kilgarriff *et al.* 2014) hosts 38 corpora from the *TenTen* family.[1] The latest (2015) edition of the English language version, *EnTenTen*, contains 15 billion words of online written English. Unlike most corpora, the web consists of a mixture of texts that have been carefully prepared and edited, along with what McEnery and Hardie (2012: 7) describe as more 'casually prepared' material, and the propensity for the latter to contain nonstandard orthographic representations can prove problematic when searching for online uses of a word or phrase of interest. Furthermore, the Web is not clearly divided by genre, with the texts returned by a search engine query, for example, constituting an undifferentiated mass of language data that is likely to require a significant amount of processing and grouping prior to analysis. Recent research indicates that organising online language in this way, for example according to register, may be challenging (Biber and Egbert 2018). The ever-changing nature of the web can also make reproducing analyses impossible, as the results of a Web search can change along with the addition and removal of online material, though downloading and archiving data can help here, particularly in cases where that data can then be shared and made available to other researchers in the future.

Of course, not all corpora fit neatly into the distinction between balanced and monitor corpora and it is even possible to combine approaches. For example, the Corpus of Contemporary American English (COCA; Davies 2009) can be described as a monitor corpus in the sense that texts are continually added to it, yet these additions are subject to a stringent sampling frame which ensures that the corpus is balanced in terms of the text varieties it contains.

Size

[1] For a list of the *TenTen* corpora (and others) hosted on *Sketch Engine*, see: https://www.sketchengine.eu/documentation/tenten-corpora/#:~:text=The%20TenTen%20Corpus%20Family%20%28TenTen%20corpora%29%20is%20a,specialized%20in%20collecting%20only%20linguistically%20valuable%20web%20content.

One of the most common questions that arises during corpus design is 'how big does the corpus need to be'? If we are following a strict sampling frame, such as that of the Brown corpus, then the size of our corpus can be determined for us from the outset. However, such cases aside, as in other aspects of corpus design there is no one-size-fits-all approach to how large a corpus should be. While corpora tend to be so large that their size would defy any plausible attempt at analysis by hand and eye alone (McEnery and Hardie 2012), there is little principled reasoning behind how large a corpus should be.

One school of thought is that the corpus should be as large as possible, with computational capacity and software speed being the only limits on the number of texts we collect (Sinclair 1991). Biber (1993), on the other hand, suggests that a corpus of one million words should be sufficient for undertaking grammatical studies, while Leech (2003) suggests that a similar number of words is sufficient for carrying out comparative work between language varieties. These are very general guidelines, though, and there is no consensus on ideal or adequate corpus size. Debate on this issue continues and 'big' has not always been assumed to be 'beautiful'; one clear advantage of building a smaller corpus is that the human analyst is likely to be able to account for a larger proportion of the uses of even the most frequent words and, in some cases, to account for every use of a word of interest. Additionally, a more manageable amount of data makes it easier for researchers to investigate the contexts in which the texts in the corpus were produced, as well as to link these insights to analytical findings (Baker 2006) (see also Chapters 5 and 6).

Determining how large our corpus should be is, however, a luxury that we are not always afforded. As noted in the previous section, in some cases it can be beneficial to adopt an 'opportunistic' approach to corpus design by collecting all texts that are available to us. The design of such corpora does not seek to address issues relating to balance or skew but, rather represent 'nothing more nor less than the data that it was possible to gather for a specific task' (McEnery and Hardie 2012: 11). For example, consider extinct languages for which a body of literature survives, such as Classical Latin, may allow us some degree of choice in corpus design. However, for other languages, such as Eteocypriot, the surviving texts are much fewer in number, and our understanding of them is limited, meaning that our choice in corpus building is limited and a corpus approach to the language is neither necessary or, perhaps, credible.

In summary, the size of the corpus we build is likely to depend, on the one hand, on the type of analysis that we want to carry out on it and, on the other, on practical considerations and limitations regarding what is possible. Whatever the size of our corpus, it is important that we engage critically with what insight it does and does not have the capacity to afford.

3. Ethics and copyright

Ethical standards and principles in corpus building are, as in other areas of linguistics, widely debated, and there is no 'gold standard' for corpus builders to follow. Some researchers argue that for ethical purposes we should draw a distinction between texts that exist within public and private domains, with texts existing within private domains, which are thus likely intended for private audiences, requiring informed consent from their authors before they can be collected and studied.

Beyond the public-private distinction, we should also consider the potential risk of the research we are carrying out to do harm to those whose texts we are collecting and analysing, weighing up the benefits of the research against the potential for harm. Such decisions are, of course, not straightforward, and nor are they to be taken lightly. In any case, when collecting texts which do not exist in the public

domain, it is good practice to anonymise those texts by removing mentions of details by which the authors could be identified in so far as that is possible. For more detailed discussions of ethical considerations in the design of written corpora, see McEnery and Hardie (2012: Chapter 3), Collins (2019: 34-37) and Hunt and Brookes (2020: 77-81).

In addition to ethical considerations, we also have to be mindful of copyright restrictions that might prohibit distribution of the texts we want to include in our corpus. It is illegal, for example, to download an entire text and then redistribute it without the permission of the copyright holder. This clearly presents a problem if we intend to make our corpus available to others at any point. McEnery and Hardie (2012: 59) discuss some of the ways in which we can address copyright issues when building a written corpus. The first is to contact the copyright holder to request permission to reproduce the text(s) in question under the terms of some specified licence. This is most feasible if one or a small number of texts are to be sampled. Alternatively, we could focus our data collection only on those texts whose owners have explicitly permitted their redistribution, for example a website which declares its content to be in the public domain or which is available under a licence permitting copying and redistribution. Restricting a corpus to such texts would, however, almost inevitably lead to a skew in the types of texts our corpus then represents. Thirdly, if we collect data without permissions necessary for redistribution, we can nevertheless share our corpus with others by hosting it on a tool which allows other researchers to run concordance queries but shows only a very limited amount of text in the output. Since it is impossible to reconstruct the original texts from the tiny snippets that such a concordance would provide, which are small enough to count as 'fair use', this 'redistribution' is unlikely to constitute a dangerous copyright violation.

Finding legal ways of sharing our corpus is useful not only to those who set out to provide a public resource, but this can also help with research ethics, as it helps to ensure the replicability of our research. Such concerns may lead us, where for legal reasons we cannot distribute a corpus, to provide a clear set of instructions for recompiling a corpus, legally for the purposes of replication. For example, if a corpus has been created from a news consolidation service which prohibits text sharing, then researchers may publish the parameters and query used to gather data from that service so that other users with access to it can, effectively, re-recreate the corpus for themselves. A final point on copyright which is important to bear in mind is that, when collecting online texts, it is important to be sensitive to potential differences in copyright and fair dealing laws across the various geographical zones represented by our data. What is legal in one jurisdiction may be illegal in another and corpus builders should be aware of that.

4. Text gathering and processing

In this section, we will consider: text collection, after which we will consider two processes commonly undertaken on texts collected for inclusion in a corpus - cleaning and annotation.

Text collection

The Internet has undoubtedly transformed the process of corpus building, as it grants instant access to an unimaginable number of downloadable texts that already exist in an electronic format. These texts do not just constitute e-language, as many texts that were originally written in the more traditional sense are now available on websites or online archives, with such resources providing much more convenient means for corpus building than their original paper forms. For example, websites like *LexisNexis* can be used to download the text from large numbers of news articles according to user-

determined criteria, while *Project Gutenberg* allows one to download copies of literary texts which can then be stored and analysed as corpora. If we decide to download texts from online sources such as these, we may have to convert them into a format that is suitable for the tool we are using. This may mean that we need to convert some texts from one format to another, e.g. from a pdf into plain UTF8 text.

If we want to collect data from a large number of websites or from websites that have a lot of pages, it can be helpful to use a website copier like *HTTrack* or *BootCat* (Baroni and Bernardini 2004), which can scrape all text, and in the former case images and hyperlinks, from user-determined webpages with impressive efficiency. Online texts can also be retrieved in a less structured manner using *BootCat*, which can compile a relatively unstructured corpus of texts, in terms of registers, by trawling the web using search-terms specified by the user.

If the texts we want to include in our corpus are not readily available in electronic form, or do exist electronically but in the form of graphics files not amenable to corpus processing, then they will either have to be keyed in by hand or, if the print quality is sufficient, scanned in using Optical Character Recognition (OCR) software. In most cases, scanning texts is more efficient than keying them in, which can be extremely time consuming. However, scanning also presents issues as OCR can struggle with texts whose pages are structured into columns. OCR is also prone to error, especially if texts are damaged or of low quality. This is a particular issue when compiling corpora of historical written texts, where the results will usually have to be corrected by hand. This task generally becomes more painstaking the larger our corpus is and the older the texts in it are. This process brings us to the next step in written corpus construction: cleaning.

Cleaning

The texts we collect for a corpus may need to be processed further before we can start analysing them as some of their features may adversely affect the accuracy of the analytical procedures we intend to carry out, as well as impinging on the corpus's representativeness in a more general sense.

The first cause for concern in this regard is the presence of so-called 'boilerplate' text. This is language that occurs within the texts we have collected but which is likely to constitute 'noise' in the context of our corpus and which gets in the way of our analysis. For example, if we download news articles from *LexisNexis*, the text files given will include labels which denote what is the 'headline', 'byline' and 'author'. Found in every text, the occurrences of these elements can accumulate quickly and thus become a problem for frequency-based corpus measures. Tools can help with this – for example, *WordSmith Tools* (Version 7 on; Scott 2016) now includes a 'Boilerplate removal' function. Rather than simply removing such elements, we may wish to use them as a form of metadata, in which case they can provide the basis for corpus annotation – discussed below. Deciding on what counts as 'boilerplate' material, however, is a subjective judgment and may depend on our research aims (Collins 2019: 40).

Another issue that can arise during the collection of online written material is the presence of duplicate texts. For example, *LexisNexis* can store multiple versions of the same news text, such as online and print versions of the same article, as separate downloadable files. Again, tools can help - version 7 of *WordSmith Tools* includes a 'duplicate text' function which ranks texts by similarity, allowing the corpus builder to check high similarity results and, if they wish, remove duplicate texts from their corpus. Like the identification of boilerplate material, though, deciding on what counts as

duplicate texts will depend on the aims of our research – some research, for example, may wish to explore small differences between online and print versions of a newspaper text.

Another consideration that arises during corpus cleaning relates to the presence of nonstandard orthography. Inconsistent orthographic representation is a characteristic feature of much user-generated online content, which can contain typos, abbreviations and spelling mistakes, while texts sampled from different geographical locations and historical periods can exhibit distinct spelling conventions even within the same language. For this reason, corpus builders who do not wish to study such variation as a linguistic phenomenon may choose to standardise the spelling in their corpora so that they can improve the accuracy of their frequency counts. Tools like *VARD* (Baron 2011) are useful for this purpose, as they can quickly scan for instances of nonstandard orthography and present these to the user, at which point the user can make a decision to either standardise spelling variants or retain them in their current form. Where standardisation occurs, the original spellings are typically retained as a matter of best practice, with corpus markup used to differentiate the standardised and original forms.

Text Encoding

To help analyse our data, there are three types of information that we may wish to encode in our corpus: metadata, textual markup and linguistic annotation. Metadata is information about the text itself. For a written text, this may include information about its author, language and date of publication. The second type of text encoding, textual markup, typically represents paralinguistic features of the texts in the corpus, for example denoting where italicisation might start and end in a stretch of writing, or telling us where an image occurs within a text. Finally, linguistic annotation can mark linguistic features we believe to be implicit in our texts, e.g. we can linguistically annotate our corpus with information about parts of speech, lemmas, grammatical structures and semantic categories, *inter alia*. Whether or not we annotate our corpus and with what types of information will depend, again, upon the kinds of analysis we want to carry out as well as the resources we have available. Resources for annotating corpora are ever more widely available. As well as online systems which provide such features as part-of-speech and semantic annotation, tools such as *#LancsBox* (Brezina et al, 2015) and *Sketch Engine* provide annotation for parts-of-speech in a range of languages. For example, *#LancsBox* provides part-of-speech annotation for over twenty languages, and lemmatisation for a subset of these. *Sketch Engine* also provides part-of-speech tagging and lemmatisation for a wide range of languages.

Metadata, textual markup and linguistic annotation are all usually encoded in corpora using *eXtensible Markup Language* (XML). The use of XML is standard not only in the annotation of corpora but also, for example, the reliable transfer of webpages and word-processor documents from one machine to another. Using XML, tags are contained within angular brackets (<tag>) which makes the tags searchable but also allows the words within the brackets (i.e. the tags themselves) to be excluded from corpus analytical procedures.

Linguistic annotation may be introduced into a corpus automatically, semi-automatically or manually. Automatic annotation is, naturally, appealing and for some tasks it can be carried out with a high degree of accuracy. For example, the CLAWS tagger (Garside *et al.* 1987) annotates texts for parts of speech with around 97% accuracy. Similarly, VARD (Baron, 2011) can generally, successfully link words in irregular spelling in a corpus to their standard form, with the standard form introduced as a linguistic annotation encoded in a suitable markup language, e.g. XML. Yet some types of written

texts are particularly likely to pose challenges to automated taggers. For example, orthographic and grammatical variation in historical texts can pose processing issues to taggers trained on contemporary and standard forms, as can texts that have the propensity to be grammatically unpredictable, like learner language and unedited and user-generated e-language. Automated taggers are also likely to struggle with texts containing more than one language, while the 'patchiness' (McEnery and Hardie 2012: 31) of taggers for languages other than English remains a challenge to the field. Regardless of the type of text we are analysing, it is advisable to manually check automatic tags to correct errors. One approach to this is to create a small, 'gold standard' subset of your corpus which can be manually checked to measure tagger performance (McEnery and Hardie 2012: 31).

Whichever approach to annotation we choose, Leech (1993) proposed a number of maxims which should be followed as closely as possible: 1) it should be possible to remove the annotations and revert back to the raw corpus; 2) it should be possible to extract the annotations themselves from the texts for storage elsewhere; 3) the annotation scheme should be based on guidelines that are available to the end user of the corpus; 4) it should be made clear how and by whom the annotation was carried out; 5) the end user should be made aware that the annotations are not infallible; 6) annotations should be based, as far as possible, on widely agreed upon and theory-neutral principles; 7) no annotation scheme has the *a priori* right to be considered as a standard; standards, where they exist, emerge through practical consensus (see also: McEnery and Wilson 2001: 33-34).

Annotation can add value to a corpus, making it easier to search in a linguistically meaningful way, and so we would encourage corpus builders to annotate their written corpora if they can envisage ways in which the resultant tags could enrich their analyses. However, annotation is not essential for corpus analysis and, since it can be a time-consuming and resource-draining process, we would caution against annotation for annotation's sake.

5. Challenges and future directions

The structure of the sections 2-4 reflects the order in which much corpus construction occurs. However, this is not hard-and-fast, and it is quite normal for corpus construction to be iterative, e.g. corpus annotations may be added after a first version of a corpus is complete. We must also note that there are some potentially important aspects of corpus construction that we have not explored in much detail here, such as ways of sharing corpus data and evaluating representativeness following construction (for a discussion, see Love 2020).

Throughout, however, corpus builders should always be guided in their choices by the purposes that they intend for the corpus they are building. Importantly, those choices and the rationale behind them must be documented in a detailed and comprehensive manner in order for others to evaluate, replicate and potentially use our corpus in the future.

Our exploration of the considerations underlying corpus design in this chapter has highlighted a number of challenges that persist with regard to the capacity of written corpora to represent the texts on which they are based. For texts derived from contexts that are characteristically multimodal in their design, such as news articles and online texts employing memes and emoji, we anticipate development in the capacity of corpora to represent visual elements in ever more sophisticated ways. Research in this area is ongoing and promising. For historical texts, there is an urgent need for ever more sophisticated means of gathering and accurately formatting historical documents for corpus analysis. Development in this area could lead to the construction of larger corpora which reach ever

farther into the past without users becoming concerned about their accuracy. Finally, the development of written corpora perhaps more than any other type brings to the fore issues relating to 'aboutness' in corpus design. Many written corpora, especially those containing news texts, are designed to represent texts 'about' particular topics. The concept of 'aboutness' in the context of corpus linguistics is rather vague yet seems to underpin, both explicitly and implicitly, much written corpus design, being employed in seemingly inconsistent ways. Designers of written corpora, and corpus linguistics in general, would thus benefit from greater theoretical and empirical engagement with the concept of aboutness and its consequences for corpus design in the future.

Further reading

Love, R. (2020). *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. London: Routledge. (This book provides a refreshingly candid account of the challenges associated of corpus design and how these can be overcome).

Collins, L. (2019). *Corpus Linguistics for Online Communication: A Guide for Research*. London: Routledge. (This book introducesthe construction and analysis of online corpora, including discussing ethical considerations of online text collection).

References

Aston, G. and Burnard, L. (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh: Edinburgh University Press.

Baker, P. (2006) *Using Corpora in Discourse Analysis*, London: Continuum.

Baker, P. (2009) 'The BE06 Corpus of British English and recent language change', *International Journal of Corpus Linguistics* 14(3): 312–37.

Biber, D., and J. Egbert (2018) *Register Variation Online*, Cambridge: Cambridge University Press.

Baron, A. (2011) *Dealing with Spelling Variation in Early Modern English Texts*, Unpublished PhD thesis, Lancaster University.
Baroni, M. and Bernardini, S. (2004) 'BootCaT: Bootstrapping corpora and terms from the web', *Proceedings of LREC 2004*.

Biber, D. (1993) 'Representativeness in corpus design', *Literary and Linguistic Computing* 8(4): 243-57.

Brezina, V., McEnery, T. and Wattam, S. (2015) 'Collocations in context: A new perspective on collocational networks', *International Journal of Corpus Linguistics* 20(2): 139-73.

Brookes, G. and Baker, P. (2021) *Obesity in the News: Language and Representation in the Press*, Cambridge: Cambridge University Press.

Collins, L. C. (2019) *Corpus Linguistics for Online Communication: A Guide for Research*, London: Routledge.

Davies, M. (2009) 'The 385+ Million Word Corpus of Contemporary American English (1990-present)', *International Journal of Corpus Linguistics* 14(2): 159–90.

Garside, R., Leech G. and Sampson, G. (eds) (1987) *The Computational Analysis of English: A Corpus based Approach*, London: Longman.

Gatto, M. (2014). *Web As Corpus: Theory and Practice*. London: Bloomsbury.

Hawtin, A. (2018) *The Written British National Corpus 2014: Design, compilation and analysis*, Unpublished PhD thesis, Lancaster University.

Hundt, M., Sand, A. and Siemund, R. (1998) 'Manual of Information to Accompany the Freiburg-LOB Corpus of British English ("FLOB")', [online], Available at: www.hit.uib.no/icame/flob/index.htm.

Hundt, M., Sand, A. and Skandera, P. (1999) 'Manual of Information to Accompany the Freiburg Brown Corpus of American English ("Frown")', [online], Available at: http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM.

Hunt, D. and Brookes, G. (2020) *Corpus, Discourse and Mental Health*, London: Bloomsbury.

Huber, M., Nissel, M. and Puga, K. (2016) *Old Bailey Corpus 2.0*. [Website], URL: http://fedora.clarin-d.uni-saarland.de/oldbailey/index.html.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013) 'The TenTen corpus family', *7th International Corpus Linguistics Conference*, pp. 125–27.

Johansson, S., Leech, G. and Goodluck, H. (1978) *Manual of information to accompany the LancasterOslo/Bergen corpus of British English, for use with digital computers*, Oslo: University of Oslo.

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

Kučera, H. and Francis, W.N. (1967) *Computational Analysis of Present-Day American English*, Providence: Brown University Press.

Leech, G. (2003) 'Modality on the Move: The English Modal Auxiliaries 1961-1992', in R. Facchinetti, M. Krug and F. Palmer (eds) *Modality in Contemporary English*, Berlin and New York: Mouton de Gruyter, pp. 223–40.

Love, R. (2020) *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*, London: Routledge.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017) 'The Spoken BNC2014: designing and building a spoken corpus of everyday conversations', *International Journal of Corpus Linguistics* 22(3): 319-44.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*, Cambridge: Cambridge University Press.

McEnery, T. and Ostler, N. (2000) 'A New Agenda for Corpus Linguistics – Working with all of the World's Languages', *Literary and Linguistic Computing* 15(4): 403-20.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics: An Introduction*, 2nd edn, Edinburgh: Edinburgh University Press.

Mishan, F. (2004) 'Authenticating Corpora for Language Learning: A Problem and its Resolution', *ETL Journal* 58(3): 219-27.

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

Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.
Widdowson, H.G. (2000) 'On the Limitations of Linguistics Applied', *Applied Linguistics* 21(1): 3-25.

Wojcik, S., Messing, S., Smith, A., Rainie, L. and Hitlin, P. (2018) *Bots in the Twittersphere: An Estimated Two-Thirds of Tweeted Links to Popular Websites are Posted by Automated Accounts not Human Beings*, Pew Research Center.