6

How to Build a Corpus

In this chapter, we will present the best practices for creating a corpus. First, we will discuss some facts that need to be considered before deciding to create a new corpus and highlight the advantages of reusing existing data whenever possible. Then, we will address various important methodological concerns for creating a corpus, in particular questions related to the size and representativeness of samples, and will explain simple methods for data sampling and coding. We will also briefly discuss the challenges posed by the creation of the spoken corpora. We will finally see that the task of creating a corpus carries with it a certain number of ethical and legal issues which must be dealt with.

6.1. Before deciding to build a corpus

The first element to check before starting to compile a new corpus is whether existing data can be used for the planned study. As we will see throughout the chapter, creating a corpus is a challenging task and presents many difficulties. It is actually not always easy to find texts available in a digitized format for all text genres, and even when such texts exist, they might not all be usable due to copyright issues. Choosing the right texts to be included in a corpus should also be the object of careful reflection, since any kind of analysis carried out on data that are not representative of the target genre (see section 6.2) could be largely invalid. When it comes to creating a reference corpus, the data collection phase is so time-consuming that it can only be tackled by a group of experts. Becoming involved in a corpus creation project individually is realistic only in the case of specialized corpora, for example, if the task is narrowed to a specific language register

or a regional variety, that is, a project of a smaller size. Even for this type of corpus, several months of work are often necessary for collecting the data, and may take even longer if the latter are enriched with linguistic annotations (see Chapter 7).

The problems are even more complex and numerous when it comes to spoken data. These need to be collected in the form of audio files which are later transcribed to become analyzable with corpus searching tools. The transcription process itself is very time-consuming and its complexity depends on the exact type of annotation that is added to the data (prosodic contours, etc.). To get an idea of the magnitude of the task, up to 15 hours of work are necessary to transcribe one hour of recording (Reppen 2010b, p. 34). Transcription also poses methodological challenges (see section 6.5), for example: how should we annotate hesitations, false starts and variations in pronunciation? How should we account for the overlaps in speech turns in conversations? In addition, the use of spoken data often requires aligning the transcription with the sound file, so as to offer users the possibility of listening to excerpts from the corpus. This sound/text alignment process requires the technical know-how which can be difficult to acquire for inexperienced researchers. For all these reasons, it is preferable to reuse existing data whenever possible.

In Chapter 5, we saw that many corpora in French have already been created, and that some of them are available free of charge to the public. Some other European languages, not only English but also German, Dutch, Spanish and others, have an even broader choice of corpora than French. So, when formulating an empirical research question, it is advisable to consider whether these resources could not be used for the study. If necessary, existing data can be supplemented with a smaller portion of new data, and thus significantly simplify the data collection phase. For example, an empirical study on the regional differences in the way questions are formulated in spoken French could reuse data collected from different spoken corpora, including France, French-speaking Switzerland, Quebec and Belgium. If the study were to be extended to other regional French varieties, for example, the French spoken in the Caribbean islands, existing data could be supplemented by samples of such variety. In Chapter 4, we also saw that comparable corpora can often be assembled from existing data. For example, Crible (2018) studied how discourse markers like *bon*, *ben* and *voilà* in spoken French and *well*, *I mean*, *you know* in spoken English are used in eight different spoken registers which vary depending on certain parameters,

such as the degree of pre-planning and whether they were dialogues or monologues. In order to be able to work with comparable multilingual corpora in each speech genre, Crible reused existing data. In English, she used a British portion of the *International Corpus of English* (ICE-GB) and in French, due to the absence of a reference corpus, she assembled a corpus from a series of existing spoken (transcribed) corpora, such as the Valibel database and the *CLAPI* corpus (see Chapter 5).

If, after research, it turns out that the existing corpora are not suitable, then the creation of a new corpus might be considered. In this case, it is essential to properly outline the research question that will be studied on the basis of new data, since the latter will have a crucial influence on the whole process, both during the data collection and the annotation phases. In the field of corpus linguistics, it is very common to hear that there are no good or bad corpora, rather there is only corpora which are more or less suitable to address a certain research question. For example, for investigating the expression of subjectivity in journalistic discourse, a corpus entirely made up of editorials would not be appropriate, since this is only a sub-section of the genre, which incidentally is more likely to contain markers of subjectivity than other sub-genres, as dispatches for instance. In this case, two scenarios are possible: either the conclusions of the study will be limited to the editorial style, or the corpus should be diversified in view of including other types of journalistic texts. The problem we have just mentioned involves a key methodological point for corpus studies, which is the representativeness of data. We will discuss this point in the next section.

6.2. Establishing the size and representativeness of data

Let us begin by repeating that there is no ideal size for a corpus, in the same way as there are no intrinsically good or bad corpora. Suffice it to say that the characteristics of a corpus may be more or less appropriate for answering a research question. As the technical capacities of computers have evolved, it has become possible to collect ever larger corpora. Currently, some corpora such as the *Google Books* corpus (see Chapter 5) and the *FrenchWeb 2012* corpus (available on *Sketch Engine*), collected from the Internet, contain several billion words. For a long time, the rule of thumb for collecting a corpus was that it should be as large as possible. The logic behind this principle was that the larger the corpus, the more likely it would contain occurrences of rare linguistic phenomena. Indeed, when the words of

a corpus are listed following their frequency order, as in Table 6.1 for the *Sciences Humaines* corpus (available on Ortolang, see Chapter 5), we can observe that the frequency of words decreases very quickly from the beginning of the list.

de (9,174)	et (3,829)	en (2,495)
la (6,405)	le (3,707)	un (2,468)
l’ (4,217)	à (3,429)	une (2,455)
des (4,123)	d’ (2,586)	du (2,202)
les (4,036)	est (2,496)	que (2,122)

Table 6.1. *List of the 15 most frequent words in the* Sciences Humaines *corpus*

Frequency distribution follows Zipf’s law (1932), according to which the most frequent word in the corpus is approximately twice as frequent as the second one, three times more frequent than the third, and so on. The word frequency indicated in Table 6.1 does not follow this decrease exactly, because in French the most frequent words take up different morphological forms. The frequency of lemmatized words is closer to the curve predicted by Zipf. In any case, word frequency declines very quickly in any corpus and many words appear only once. For example, the 100th most frequent word in the *Sciences Humaines* corpus, which is the word *été*, only appears 195 times, whereas the 1,000th word, *devenir*, appears 21 times. Out of the 15,617 different words in the corpus, from the 8,355th position onwards, words only have one occurrence, meaning that almost half of the words in the corpus only appear once. Technically, these are called “hapax words” or *hapax legomena* which in Greek means “mentioned once”. We can infer that, due to this distribution, the study of rare words requires the use of large corpora in order to be able to analyze multiple occurrences.

However, more recently, some researchers have defended the idea that maximum size should not always be the goal in the creation of a corpus, since smaller-sized corpora may prove to be adequate for many research questions which do not involve rare words, as we will see in this chapter. In fact, a large corpus is not always suitable for addressing all kinds of research questions. The question of the optimal size for a corpus primarily depends on the nature of the linguistic phenomenon to be studied. The more frequent a linguistic phenomenon, the better it can be studied on the basis of a small

corpus. Rarer linguistic phenomena, on the other hand, require larger corpora. This question is also related to the degree of generalization targeted. A corpus representing a specific genre can be relatively small, whereas general corpora need to be much larger.

As we discussed in Chapter 1, a corpus does not simply represent a collection of randomly chosen texts. In fact, a corpus is a collection of texts or recordings specifically chosen in order to be representative of a language, of a certain register or even a language variety. The question of representativeness is therefore essential so that a corpus can be used for answering a research question. In order to fully understand what this notion represents, we will draw an analogy with opinion polls. Let us imagine that we wish to find out which candidate is more likely to be elected in the next presidential elections. In order to find out, it is not possible to ask all the citizens who they intend to vote for. It is therefore necessary to prepare a sample of the population of a more modest size, to whom it might be possible to ask such a question. Later, the results obtained on the basis of this sample can be extrapolated to the entire population. But in order for this technique to work, it is crucial to carefully choose the sample of respondents, in such a way that it represents the whole population. For example, if the sample chosen includes 500 students met at the exit of a university building, the sample obtained will most likely not correspond to the actual result of the election, since this sample is not representative. In fact, students represent only a small portion of the population. In order for the sample to be representative, it should also include people with other types of occupations, different age ranges and from different regions. The same applies to the compilation of a corpus. In order to be a representative, a reference corpus should contain a balanced set of samples covering the main stylistic genres, both in the spoken and written modes. The main issue is to determine the criteria according to which it is advisable to classify the elements included in the corpus to ensure its representativeness. We can be sure about one thing: these criteria should not be related to the linguistic content of the samples, but rather to a classification made on the basis of external criteria, such as text genres and language registers. For example, it would be rather inappropriate to try to study the production of speech acts in the legal context by choosing a corpus exclusively based on a number of performative verbs such as *demand*, *condemn*, *order* that it contains, since this criterion would influence the results found in the analysis afterwards. However, this study would require the assembly of a corpus which tangibly represents the legal language, such as court decisions, because these writings

properly match the targeted field of study, that is, the legal language. To sum up, Biber (1993) argued that a selection based on linguistic criteria specifically related to the content of the corpora would drive the analysis work in the direction of a circular path. Indeed, corpora should be used for analyzing the words and the linguistic structures they contain, among other things. If these parameters have been predetermined during the corpus assembling phase, this analysis no longer makes sense.

In the case of reference corpora, sample distribution between different genres is a complex problem in itself, due to the lack of an existing typology of spoken and written genres that is unanimously accepted. To simplify, let us say that written corpora should contain both public texts (published works) and private ones (letters, emails, etc.), collected from different fields such as the press, the sciences or the literature. The spoken section of a corpus should reconcile a variety of choices. It should include both planned and spontaneous spoken speech, monologues and dialogues, drawn from contexts with various degrees of formality. Very often, the creators of new corpora solve the problem of representativeness by following the criteria used in existing reference corpora, such as the *British National Corpus*, a pioneer of the genre.

In cases where researchers need to compile specialized corpora, the question of representativeness is posed a little differently. To continue developing the analogy with polls, if the goal is to know who students will vote for in the presidential elections, it will be enough to interview a sample of students, since such a sample is representative of that population. In the same way, the question of the representativeness of specialized corpora is clearly simplified, because this can be achieved by choosing texts or recordings belonging to a specific genre. However, we should keep in mind that there may be sub-genres within a genre, such as novels, short stories or children's stories within the literary genre, and that these may vary from each other.

Even when working within a text genre, we should aim to diversify its sources as much as possible. For a literary corpus, for example, works from different authors should be included.

From a lexical perspective, the representativeness of data in specialized corpora, such as corpora devoted to newspaper or legal articles, can be measured using the concept of saturation (Belica 1996). This notion means

that a certain lexical trait varies very little throughout the corpus. In order to measure saturation, the corpus should be divided into several segments of equal size. A corpus is saturated at the lexical level when the addition of a new segment results in approximately the same number of new words as the previous segment. The usefulness of this measure is nonetheless limited, since it only provides information about the lexical diversity of a corpus, but not about other domains of language.

As a matter of fact, the representativeness of a corpus cannot be ascertained once and for all. In the case of closed corpora (see Chapter 1), data aging implies that they are no longer representative of the most recent developments in the language. In the case of monitor corpora, while the new data added at regular intervals may improve their representativeness, they nonetheless pose balancing problems between the different portions of the corpus, an aspect we will discuss in the following section.

In summary, McEnery *et al.* (2005) are totally right when they affirm that the representativeness of a corpus is more a profession of faith than a scientific reality. From a factual point of view, representativeness cannot be taken for granted. What should really be kept in mind though is the need to build a corpus that best reflects the linguistic style to be studied based on available data, in order to be able to draw appropriate conclusions.

6.3. Choosing language samples

To achieve the representativeness aim discussed above, a corpus should include a sampling of different types of texts or recordings. Unless we are working on a very specific corpus like the Bible or the complete works of an author, most of the time, it is actually impossible to include all the texts or all the recordings belonging to the genre to be studied in a corpus. This is why it is necessary to prepare samples which, once assembled, can work as a representative sub-section of the genre to be studied. The preparation of samples to be included in the corpus poses two important methodological questions: on the one hand, the appropriate size for each sample, and, on the other hand, how to balance the portions of the corpus in such a way that the result is truly representative of the genre.

In order to understand the difficulties of corpus balancing, we will give an example. To be representative, a spoken French corpus should include

speakers from different regions, different ages, both male and female. If 95% of the corpus is made up of Parisian speakers aged between 20 and 30 years, such a corpus will not be a representative sample of the population, even if this means including the other speakers among the remaining 5%. While it is true that a sample is expected to bring together a rather restricted version of the overall population, it should also reproduce its main features so that the results obtained on this sample can be extended to the entire population. A corpus of spoken French should therefore include a similar proportion of male/female speakers, from different age groups and different regions. Many other criteria could be included in this selection, such as the socio-economic level of the participants, for instance. As with the question of corpus size, the balancing criterion largely depends on the question the corpus will help to study. In general, a corpus should not be used for establishing a contrast between elements which have not been balanced during the corpus compilation phase. For example, a corpus created for studying the pronunciation of vowels in Paris and Marseille and which has not been compiled representing a balanced sample of different age groups cannot be used carelessly for studying the evolution in the pronunciation of vowels between generations.

In the case of written language general corpora, it is important for the chosen samples to represent different genres, including both published and unpublished texts. In the case of the *British National Corpus* (Aston and Burnard 1998), an English reference corpus, the written texts included were chosen according to three criteria:

– the field, that is, the topic explored in the text;

– the time when the text was produced;

– the distribution mode, depending on whether it was a book, a newspaper or an unpublished text.

The spoken samples were chosen on the basis of demographic criteria such as age, gender, geographic region and social class, as well as contextual criteria. However, the creators of large corpora have agreed that it is very difficult to fulfill all the criteria to achieve a perfectly balanced corpus. One of the major problems is the difficulty of incorporating new data, an aspect which tends to create bias around the choice in favor of more readily available data. Such difficulty is largely due to copyright issues, which we will address in section 6.6. This issue prevents the inclusion of recently

published texts in a corpus that have not yet fallen into the public domain. In addition, published texts are generally more easily accessible than unpublished texts, such as emails or personal letters. Finally, texts published on the Internet are much easier to access than texts published on paper. These differences inevitably induce a certain bias towards specific text categories. In the end, balancing a corpus is never a perfect task. As Nelson (2010, p. 60) pointed out, the end result of a corpus is always a compromise between the desires of its creators and the data that can be obtained.

In concrete terms, balancing the portions of a corpus can be achieved by defining a sampling frame which delimits the population to be sampled and lists its relevant properties. Each of these properties should then be mirrored by the corpus sample proportionally to its prevalence in the population. For example, in order to create a corpus of French, speakers from different regions should be included in the sample. Therefore, geographic region becomes a relevant trait for the sampling frame of a spoken French corpus. The number of French speakers to be included for each region can be determined proportionally to the number of French speakers living in the different regions sampled. However, in many cases, these proportions are difficult to determine accurately. For example, it is difficult to determine exactly what proportion of the texts published every year belong to the fiction genre and how many are non-fiction. In this case, obtaining the exact figures is undoubtedly possible, but highly complex. The problem becomes even more challenging for the categories of unpublished texts, for which there are no existing figures. In these cases, the classification is often based on common sense or on the pragmatism of the corpus designer, depending on the importance of subcategories for addressing the questions that the corpus is supposed to help study.

Now, let us move on to the question of which samples to include in the corpus. The first important question is how these samples should be chosen. A first technique involves choosing the samples completely at random, the idea being that out of the total number of samples in the corpus, the most frequent characteristics will eventually stand out on their own. However, we cannot take for granted that this method yields a balanced sampling, especially in the case of a small corpus. As previously mentioned, a better method might be to define a sampling frame and to divide the samples to be collected depending on the important properties of such a frame. For example, if the sampling frame for a corpus of French spoken in Switzerland includes different criteria such as gender, age, socio-economic level or place

of residence, an equivalent number of samples should be chosen to match each selected criterion. Within each criterion (e.g. 20- to 30-year-old middle-class men living in the canton of Geneva), participants can be chosen at random. According to Biber (1993), this technique, which is called stratified random sampling, never gives less representative results than purely random sampling, and most of the time, its results are much more representative than a random selection.

The next question to consider is related to the number of samples required in the corpus, as well as the ideal size for each sample. Once again, the answers to these questions depend on the type of corpus the researcher has in mind. The more generic the corpus, the more samples will be needed, whereas for a more specialized corpus, fewer samples are necessary in order to provide representative data. On the basis of corpus studies on the differences between genres and between language registers, Biber (1993) argues that in most cases, 10 samples of 20,000 words per genre are enough to obtain representative samples for each genre. Indeed, from this size onwards, the new occurrences found by incorporating additional samples become smaller in number. In the case of a reference corpus like the *British National Corpus*, 40,000 word samples were retained (Nelson 2010, p. 59).

Finally, another important question concerns sampling units. Should we include whole texts or only excerpts, or even isolated sentences? The answer to this question often depends on the accessibility of data. On the one hand, it is preferable to create a corpus including language samples which represent a coherent whole, rather than isolated sentences. However, this is not always possible due to copyright reasons. The correct size of samples also depends on the type of text considered. For example, if the goal is to reach a sample of 200,000 words per text genre, this number of words can almost be instantly reached by including one or two entire books in the sample. In this case, it would be a better idea to choose excerpts (e.g. chapters) from different books, instead of a longer portion of a single book. For other types of text such as letters, text messages, etc., the units are so small that it makes no sense to not fully include them. In all cases, it is important to systematically avoid including the same portions of text, for example, always the beginnings or the endings. Indeed, Stubbs (1996) observed that there are very few linguistic features which remain constant in a text. In order to observe all linguistic phenomena, it is therefore necessary to modify the text portion (beginning, middle, ending) that is included in the corpus.

6.4. Preparing and coding corpus files

In order to include language samples in a corpus, first we have to obtain them. In the case of spoken corpora, data acquisition first requires them to be transcribed. This is a very complex process, and we will discuss it in detail in the next section. In the case of written corpora, the situation is not always simple, either. The most favorable scenario is clearly the one in which data are readily available in digital format, which is progressively becoming more frequent, especially when it comes to data gathered from the Internet. However, retrieving text from digital formats may have varying degrees of difficulty depending on the original format; for example, it can be very difficult to isolate the texts of different articles in a journal page formatted as a PDF document, or even impossible when the PDF file includes text images. So, even when we are working with data in a digital format, the task of converting the original format into a usable format, based on the text, is still necessary, as we will see below. In some contexts, however, written data are not available in digital format. In this case, we can either work with printed texts available on paper or with handwritten texts, such as student essays or private letters. Printed texts can be scanned and then processed thanks to optical character recognition (OCR) software, but these always require a manual check made by a human in order to provide a completely reliable result. There might be a high number of errors if the original print is of a poor quality. Finally, for handwritten data, there is no solution other than to manually type it on the computer. Data transcription also raises many questions related to the way in which some of their original features might be preserved. For example, student essays often contain spelling mistakes, which should be left untouched, since they can be very informative for many research questions. But in this case, a version without misspellings should also be included so that the words can be found by a concordancer. In Chapter 7, we will see that errors can be systematically annotated in a corpus, in the same way as syntactic or semantic information is provided.

No matter the way of acquiring data, an important point is to save the corpus files into a format which can then be used by a concordancer. As we saw in Chapter 5, most concordancers (like AntConc) only read files in text format, which can include texts tagged in XML format. Therefore, all newly created files for a corpus should be directly saved into text format. Files which have already been scanned are rarely saved in this format, since this format does not make it possible to include formatting marks, and this makes documents difficult to read. In the case of corpus studies, this is not a

problem, however, since the files will not be read by humans, but processed by a concordancer. Files processed with OCR software or word processing tools are often saved in proprietary formats (such as Microsoft Word, for example, DOC or RTF). Files saved in these formats can be easily transformed into text files thanks to specific options such as the "save as" command found in word processors.

Web pages are available in HTML format. This format contains many formatting marks in the form of tags, which are interpreted for creating the various graphic effects which are necessary for a browser to display a web page. These later become visible when a file is opened with an editor in pure text format. Despite their lack of linguistic relevance, these tags are interpreted as textual elements by concordancers.

In order to avoid this problem, software should be used for retrieving text from these files and eliminating unnecessary markings (HTML tags). Word processing software often include this feature, simply by choosing the "text format" option from the "save as" command. The only problem with this option is that it is necessary to open every file one after the other in the word processor so as to perform the operation. This might eventually become a problem with a corpus, including thousands of different files. An alternative solution is to use the AntFile Converter, which is a file conversion software developed by Laurence Antony, the creator of AntConc (see the URL at the end of this book for AntFile Converter). This software can be downloaded for free and used for converting any number of XML, HTML or sometimes even PDF files into text format.

The *Sketch Engine* corpus creation and management platform, discussed in Chapter 5, also automatically transforms the format of files downloaded from the Web. The advantage of this platform is that it offers the possibility of automatically downloading large amounts of data from the Internet in a single operation (web crawling). The corpus created in this way can be directly analyzed using the tools provided by the platform, for example, for retrieving concordances, word lists or keywords. In its recent versions, the WordSmith concordancer also offers a similar function. This type of tool has made the collection of web-based corpora extremely easy. We should nonetheless bear in mind that the texts found on the Internet are of a highly variable quality and are not representative of the whole language.

If the corpus has been created manually rather than through the use of a platform enabling automatic data download, a practical but important question concerns the number of files that should be created. More specifically, should we create a single file for the whole corpus? Or should we create one file per corpus sub-section? What about a file per text included in the corpus? In general, it is preferable to store every language sample in a separate file. In this way, it is later easier to combine them in different ways for creating sub-corpora, rather than having to retrieve text portions from a larger file. For example, if we collect data on the language used by young people in France, we might then want to compare data depending on different criteria such as gender, geographic region or age group. This comparison can be done in a relatively simple way by grouping all the men's files and all the women's files or, for the same purpose, all the Paris files and all of the Marseille files. But if all men are included in a single file and women in another, then the geographic comparison data needs to be reprocessed.

In order to be able to easily group the files into sub-corpora, it is necessary to represent the features of each sample in an easily accessible manner. A practical solution is to code these characteristics directly onto the file names: this is why these names are another important element that should be taken into account when creating the corpus. For example, it is possible to name files only by using a number, each representing a criterion used when compiling the corpus. Going back to the example of young speakers, one possibility would be to identify all the files from Paris with the number "1", those from Marseille with the number "2", etc. Then, the second digit could be used for coding gender, "1" for women and "2" for men, then the third reference could be for coding the age group, for example, "1" for 16- to 19-year-olds, "2" for 19- to 22-year-olds, etc. Finally, several digits can be used for coding the participant's number. Following this procedure, the sample corresponding to the first 18-year-old male participant from Marseille registered in the corpus would be saved in a file called "221001.txt". The disadvantage of this method is that the coding is opaque for a user who does not have a precise vision of the system used.

A more transparent way to achieve the same result is to use abbreviations. For example, the same file could be coded using abbreviations such as *Mar* for Marseille, *h* for men (*homme*), *ado* for the 16- to 19-year-old group, which would result in a file called "mar_h_ado_001.txt", if we use the underscore symbol as a separator for the abbreviations. If this system is used, abbreviations should be kept short in order to avoid generating

excessively long file names, which might not be readable. We should also avoid inserting spaces or other punctuation marks, since these could interfere with the programs used for opening the files on different platforms (typically concordancers). Finally, if word abbreviations are used, it is desirable that each abbreviation of a category contains the same number of characters (e.g. three letters for all the names of cities), in order to make reading in columns of lists of files easier.

We have pointed out that corpus files should contain plain text, in order to facilitate data analysis. However, for a corpus file to be used as a sample representing a certain type of language, metalinguistic information (which is not part of the text or of the dialogue) should be accessible to the researchers who will analyze it. For example, this type of information includes the date of a newspaper article, the place where a conversation was recorded or the characteristics of the speakers taking part in the dialogue. This "piece of extra information concerning the data" included in the corpus is what we call metadata.

For this information to be made available for future users of the corpus without separating it from the rest of each sample portion, it should be possible to include it in the files, but in such a way that it is not taken into account by a concordancer when counting the words of the corpus. A possible solution could be to insert these marking elements inside tags, something that the concordancer will be able to ignore. Most often, these tags are delimited by chevrons (the less-than and greater-than signs < >).

In this way, the metadata of a corpus sample can be added at the beginning of each document as follows:

<texttype: newspaper article>
<publication: *Le Monde*>
<author: Jean Dupont>
<date: 1 April, 2019>
<subject: April Fool's Day>

In the AntConc concordancer, discussed in Chapter 5, it is possible to inform the program about the existence of tags and not to consider the information they contain.

In some cases, the abundance of metadata requires the use of a precise syntax for tags, based on the conventions of computer languages for XML or SGML coding. The conventions used can be explained in the corpus documentation or may follow a more widely recognized standard. Indeed, some sophisticated marking formats have already been developed for corpus data. One of the best known is the TEI format (Text Encoding Initiative), which makes it possible to encode many markings in a standardized way and make corpora sharing easier. Large reference corpora such as the *British National Corpus* are tagged following the TEI conventions. Without going into details, a TEI-tagged document always contains two types of elements:

– the header;

– the body of the text.

And these two elements are respectively made up of other elements. The header section contains metadata, a description of the file, the encoding, the text profile, mainly the language, the context or participants, and even a history of its revisions. All these elements, except for the file's description, are optional. The body of the text mainly contains tags, which are destined to delimit text units such as paragraphs or even sentences. Following XML conventions, TEI tags always begin with chevrons < > and close with </ >. As we will see in Chapter 7, TEI tagging can also be used for making more detailed annotations than dividing the text into sentences.

In addition to indicating the metadata by means of tags inside each file, it is also very useful to provide a summary table in the corpus documentation, as illustrated in the simplified Table 6.2. This type of table gives users an idea of the contents of the corpus at a simple glance and helps them to quickly choose those files which are relevant to their concerns, without having to open them all one by one.

File	Gender	Age	Residence	Context	Topic
113001.txt	Male	28	Neuchâtel	At home	Retells a memory
224002.txt	Female	32	Geneva	At work	Talks about her work
212003.txt	Female	25	Martigny	At a coffee shop	Retells a memory

Table 6.2. *Example of a table summarizing corpora metadata*

6.5. Recording and transcribing spoken data

The collection of spoken corpora poses certain additional challenges compared to written corpora. One of the main difficulties stems from the need to transform spoken data into a written format. As we have seen in Chapter 5, for corpus data to be analyzed, they should be in written format, since concordancers cannot search for words or expressions in audio files. In this section, we will briefly discuss some of the problems related to the representation of spoken data, as well as some possible solutions to sort them out.

The first step we can take to work with spoken corpora concerns the mode of acquiring data. Spoken data need to be recorded and the recording process itself requires special preparation. For data to be as representative and informative as possible, it is essential to properly define the research questions that these data will help answer well in advance. In particular, these questions will determine the type of interactions that should be recorded, the constraints regarding the context as well as the information contained in the transcript. If, for example, the aim of a spoken corpus is to study the lexical specificities of a language variety, the prosodic information contained in the interactions will be of little use. If, on the other hand, the research question concerns information structure in discourse, more specifically the introduction of new and given information in different spoken genres, then prosody will play an important role in studying the interface with the utterance structure and therefore requires a transcription.

An important point to establish before carrying out the recordings is the nature and the amount of contextual information that will need to be added to the transcripts. Spoken conversations are naturally more ambiguous and less precise than written communication, since speakers can use the immediate context to make themselves understood. Audio recordings make it almost impossible to grasp this type of information, which later have to be added to the corpus so that the interactions can be understood and analyzed by the experts who will listen to these recordings. In the case of audiovisual recordings, a larger share of contextual information will be captured and should not be explicitly added to the corpus (although some kind of codification may later be required to perform specific analyses).

Due to the difficulty of collecting and transcribing spoken data, the question of the amount of data needed for creating a corpus is even more acute than for written data. But in this case too, there is no ideal size for a spoken corpus. The amount of data required for studying a certain phenomenon primarily depends on how frequently it occurs. For example, Adolphs and Knight (2010, p. 41) have estimated that one hour of recorded conversation corresponds to approximately 10,000 words in the transcript.

As we have already mentioned, providing metadata details is particularly important in the case of spoken corpora. Among other things, the metadata appearing in the header should offer information about the main features of the participants: degree of relatedness (parents, friends, colleagues, strangers, etc.), the context in which the conversation took place, the manner in which the recording was captured, etc. The importance of this information depends on the questions that the corpus is expected to answer. For studies on how interpersonal relationships influence interactions, it will be necessary to have as much information as possible about the degree of relatedness between participants, whereas for studies on the use of discourse markers such as *bon* or *ben*, this type of information is not so relevant.

In the same way, contextual information may be added to the statements inside the transcripts. Let us insist on the fact that the context of an interaction is so rich that it would be illusory to try to account for all the aspects involved in a transcript. Choices will have to be made depending on the importance of this information for the research question. At least, the transcripts should contain enough contextual information for the meaning of the utterances to be reconstructed if this became necessary in the absence of context. For example, if a person passes by and this event invites a comment from the participants, this piece of information should be mentioned in the transcript, indicated between tags so as not to be confused with the transcription itself.

Finally, the last difficulty related to transcription that we will mention concerns the presentation of the transcripts. In a dialogue, the participants do not always speak one after the other as it happens in the dialogues of a novel or a play. On the contrary, there are many overlaps between speaking turns, as well as pauses. The analysis of overlaps and pauses can be important for certain studies, so the question arises on how to best account for these phenomena. If a transcript is presented in a purely linear fashion, one intervention above the other, valuable information might be lost. This is why

other types of presentation are often used. For example, in the CLAPI corpus, the contributions from the different participants are presented one below the other. Overlaps are indicated by green square brackets, making it easy to see where and when they occur, as shown in Figure 6.1. Numbers in light blue indicate the presence of pauses and their duration in seconds.

```
JEA bah c'est (inaud.) tout façon
JUL c'est mignon ouais/
JEA [ah oui]
JUL [nan mais][la (inaud.) c'est quand même différent en france tu vois]
LAU            [non mais linköping lin- linköping c'est                ]
    plus p`tit que besançon j` veux dire
(1.1)
JEA j` veux dire b`san::çon
JUL ((rire))
CLA < ((en riant)) huhuhuhum
```

Figure 6.1. *Example of a CLAPI corpus transcription. For a color version of this figure, see www.iste.co.uk/zufferery/corpus.zip*

Yet another solution is to represent the words from each participant in a separate column and to show the overlaps on the same line.

In summary, the transcription of spoken data requires many decisions to be made concerning the nature and the amount of information to be added, not only to the dialogues themselves, but also on how to communicate such information on the files and visually. These decisions should be made even before the data collection process begins, since an important portion of contextual information could be lost if it is not recorded during the interactions.

6.6. Ethical and legal issues

Creating a corpus involves using (or even sharing with other researchers) language samples produced by third parties. Those persons having contributed to a corpus through their language productions have rights that need to be respected. In the case of a spoken corpus in particular, it is essential for participants to know that they are being recorded and that their data will later be used for linguistic analyses. For this, the creators of a corpus must hand a document to their future participants clearly explaining who the data will be accessible to and how it will be used. The participants can then freely decide whether or not to sign a form stating their consent to take part or not in the study. However, such consent to participate does not

suffice to share the data with other people afterwards unless this usage has been explicitly mentioned in the form. In fact, a participant may agree with the idea of being recorded by a researcher and then having such data used for research, but not necessarily agree with having their data being shared with a large number of people, perhaps even with web-free access. In order to be able to share corpus data, it is imperative to ask participants both for their authorization to use and to distribute the data before collecting them. If a participant later refuses to have their data distributed, removing them from the corpus may pose many difficulties. In the case of dialogues, all of the data of the events that the participant was involved in will have to be removed.

The right to anonymity of the persons mentioned in the corpus represents another important ethical problem. Often, people interacting in recorded conversations refer to third parties by naming them. These people did not provide their consent to being talked about in public documents, so their names should be removed before publishing such data. This anonymization process is not always easy, however. For example, McEnery and Hardie (2012) have mentioned several cases drawn from the *British National Corpus* in which the people in question were very easily identifiable even after their names had been deleted, since the context was precise enough to be able to find them on the Internet. For instance, this is the case of conversations in which persons with a specific role in the village, such as the doctor or the clergyman, are mentioned. This situation is particularly problematic when references to people include degrading criticism or reveal their illicit activities. In these situations, it is necessary to delete parts of the conversation, or even the entire conversation, to protect the rights of the persons concerned.

In the case of written corpora, the situation is simpler, especially when it comes to published data. It is reasonable to think that the public figures mentioned in the articles agree to waive their right to anonymity. The responsibility of the corpus compiler is involved when it comes to texts with potentially defamatory content. In the case of articles found on the Internet, in particular, source verification is necessary before indiscriminately including texts collected automatically, following the web crawling processes described earlier in this chapter. In general, we should also be aware of the fact that distributing a corpus implicitly amounts to disseminating the ideas contained inside its texts. In some cases, this may pose ethical problems for researchers. For example, Baker and Vessey

(2018) have compiled a corpus of propaganda by extremist groups. Distributing such a corpus clearly invites the question of whether the ideas it contains should be propagated on the Internet or not. The authors were also unable to access a portion of the propaganda journals published by certain terrorist organizations. In the United Kingdom, the possession of this type of material is illegal. This example brings us to the second aspect considered in this section, namely the legal issues involved in the creation and distribution of corpora.

Written corpora containing published texts are confronted with copyright issues. While it is true that laws differ from country to country, it is not possible to distribute the texts of an author during his/her lifetime without demanding some type of compensation. In France, this period is valid until 70 years after the author's death. However, contrary to what many people think, data accessible on the Internet are also subject to copyright. Their use can be softened, providing that they are accompanied by sufficiently permissive user licenses, such as the Creative Common license which concerns the contents of the collaborative encyclopedia Wikipedia. Due to copyright restrictions, in the case of less permissive licenses, corpora creators encounter many restrictions for including data. There are several possible strategies for properly addressing the copyright problem.

First, we can limit our choice to works that have fallen into the public domain and/or coming from websites where data have been declared free of rights. This solution is the safest one from a legal point of view, though it is not the most satisfactory one from a linguistic point of view. As a matter of fact, this selection method hinders the collection of data that truly mirror contemporary language and certain stylistic genres which are poorly represented on the Internet.

A second solution would be to negotiate the right to use data with their owners. This solution can be realistic when creating a corpus drawn from a limited number of sources. Rights holders often agree to authorize a single researcher to use a reasonable amount of their data for research, but this type of corpus often cannot be later redistributed. This limitation poses a problem for research replicability, which is an important scientific element in order to grant its validity.

Another way of dealing with the copyright problem during corpus distribution would be to allow users to search for concordances in the corpus, but not to visualize it in its entirety. This is the case for many corpora that are only available via an online consultation interface. These interfaces only enable occurrence searches for words or expressions within a certain context. This solution effectively preserves copyright, since the works remain inaccessible. For users, these interfaces make it possible to answer a certain number of research questions related to the lexicon.

However, they are unsuitable for research questions that require data processing, for example, some kind of annotation or those that have to take into account a large context, in order to identify certain linguistic phenomena such as speech acts or discursive phenomena.

6.7. Conclusion

In this chapter, we have discussed the main elements to consider when creating a new corpus. For a start, we mentioned that corpus creation is a long and complicated process. This is why reusing the already existing data should be prioritized as far as possible. Then, we saw that the important methodological trait to be respected when creating a corpus is data-representativeness. The latter can only be defined in relation to a specific research question. The representativeness of a corpus also depends on its balance and the choice of samples it contains. We also introduced some basic principles regarding sample collection and balancing. We then addressed some concrete problems, related to data coding and transcription into a corpus, and concluded that these questions needed to be resolved before starting the data collection phase. Finally, we saw that the creation of a corpus poses several ethical and legal questions which should be carefully considered, since distributing data amounts to disseminating information that belongs to and concerns third parties, whose rights must be respected.

6.8. Revision questions and answer key

6.8.1. *Questions*

1) What types of data should be collected to conduct a representative study of how young people use the discourse marker *genre* in French?

2) How could we balance the different parts of a corpus aiming to study the French literature of the 18th Century?

3) What are the main questions to consider when choosing the samples to be included in a corpus?

4) Using the *Sketch Engine*, choose five keywords in order to create a corpus on the *French cinema*. What are the characteristics of the corpus thus created? Which are these keywords?

5) What transcription information would it be important to add to a corpus of spoken conversations to study the language of the suburbs in France?

6) What are the ethical issues to consider in the following cases:

a) a collection of texts produced in class by children;

b) a recording of spontaneous conversations of a group of friends at a bar;

c) a recording of a teacher's course for a spoken corpus.

7) Which of the following actions do you find problematic from a copyright perspective:

a) using a digital version of the novel series *Harry Potter* for compiling a corpus stored exclusively on your computer;

b) sharing this corpus with your partners as part of a corpus linguistics course in order to do joint homework;

c) distributing this corpus on the Internet;

d) including an entire chapter drawn from this corpus in a publication with the aim of illustrating certain linguistic phenomena that you have annotated.

6.8.2. *Answer key*

1) In order to have representative data for this research question, the corpus chosen should evidently contain language produced by young speakers. This concept would need to be clarified to be operational, for example, by deciding to include an age group ranging from 15 to 25 years.

This study does not specify a geographic region. One way of delimiting research for such a project would be to compare young people living in large cities in four French-speaking regions from different countries, for example, Paris, Brussels, Geneva and Quebec. The young speakers included in the corpus should proportionally represent the two genders and have different socio-economic profiles. Finally, the corpus should contain young speakers recorded under similar conditions in order to avoid any context-related bias. Another study could aim to compare these uses across different speech styles, which should then be represented in the corpus in a balanced manner.

2) This research question is fraught with different constraints. First, the corpus should contain French literature, which restricts the literary subject to works written in original French, rather than translations. It should also span a specific period, which could be defined, for example, as works published between 1800 and 1900. The difficult point to assemble this corpus relates to the way of balancing its content between different literary genres. It should therefore contain novels, short stories, plays and poetry. Let us assume that the corpus targets a size of 200,000 words per genre, in order to have a representative sample of each of them. Since novels and plays are long texts, it would be a good idea to include excerpts (e.g. a chapter or an act) from many different works, rather than two or three texts in their entirety. Conversely, since poems are a very short genre and more marginal in terms of the amount of texts published, a possible decision would be to limit the share of poems to a smaller percentage of the corpus, for example, to limit poetry to 50,000 words.

3) The first question to ask is whether a sample is representative of the genre it embodies in the corpus. For example, before deciding to include an interview with a young Brussels resident in a corpus of French spoken in Belgium, we should make sure that this person has not recently moved to Belgium from another country, and that they properly reflect the linguistic specificities that the corpus is supposed to embody. The second important question concerns the size of the sample that will be included in the corpus. As we recalled above, it is not always optimal to include entire texts in a corpus when these are very long. Depending on the target size for each genre, it is necessary to determine the appropriate size for each sample. A third question to consider concerns the way in which the samples are acquired, depending on whether these are digitized texts, texts to be scanned or transcribed. Besides, it is also necessary to determine which metadata will

be associated with each corpus sample. All of these decisions need to be made based on the research question being considered. A final and very important point is to ensure that copyright is respected, either because the text is copyright-free or because the author has provided written consent granting permission to include their text in the corpus. Finally, we should make sure that the sample is acceptable from an ethical point of view, in the sense that it respects the right to anonymity of the person involved and that it does not contain inappropriate content.

4) By creating a corpus with keywords such as *cinéma*, *films* and *acteurs* and using the default parameters offered on the site, *Sketch Engine* produces a corpus of 90,441 word occurrences, including 2,680 word types retrieved from 35 different pages. The most frequent content word is *cinéma*, at the 20th frequency rank. The keywords in the corpus include proper nouns such as *Edison*, *Funès*, *Fernandel*, *Gabin* and *Reynaud* and also content words like *cinéphile*, *cinéma* and *crédits*. Collocations include elements like *cinéma français*, *cinéma muet*, *art dramatique*, *histoire du cinéma*, *grand écran*, *carrière cinématographique*, *mise en scène*, *film français*, *actrice américaine*, etc. These collocations make perfect sense in view of the search terms used for creating the corpus.

5) Transcription information should include both metadata and indications inside the transcripts themselves. The metadata of such a corpus should at least include information of where the recording took place, its date, the context in which it happened, the conversation topic, the number of participants, the gender of the person recorded, his/her age and profession. Within the transcripts, it might be useful to include an annotation of the non-standard words used, for example, in verlan, with their equivalent standard so that they can be found in a concordancer. An indication of pauses, overlaps and certain prosodic phenomena may also be useful.

6) a) Above all, a **collection of texts produced in class by children for assembling a corpus** requires protecting the children's right to anonymity. No element in the corpus should make it possible to identify any participant. Depending on the nature of the texts, it is necessary for the content to exclude any element making it possible to identify any other person.

b) In the case of a **recording of spontaneous conversations of a group of friends at the bar**, everyone involved should be notified that the conversation is being recorded and that he/she agrees. Depending on the

purpose of the corpus, this agreement should also consider data sharing with third parties. Then, depending on the conversation topics, it would be necessary to ensure that the content is neither defamatory nor offensive, and that it does not enable third party identification.

c) Finally, in case we decide to **record a lesson from one of the professors for a spoken corpus**, we should make sure that the professor has been informed about the recording and has given his/her consent, both for the use and for the possible sharing of the data.

7) a) Using a digital version of the *Harry Potter* series for assembling a corpus to be stored only on your computer and searching for elements in the text is not problematic *a priori*, insofar as this digital version has been legally acquired, for example, by buying an e-book from an online bookstore.

b) However, sharing this corpus with your classmates within the framework of a corpus linguistics course with the aim of carrying out joint homework is a bit more delicate an issue, because the fact of buying a book does not entitle you to duplicate it or to transmit it free of charge to others. This practice may, however, be considered as a tolerated use of the material if the aim is to carry out joint homework on the data, which are not being used in any other way.

c) Distributing this corpus on the Internet is completely illegal under copyright rules which can be enforced for decades after the author's death (70 years in France). In this case, the *Harry Potter* series will still be protected for many years, and any form of distribution is currently prohibited.

d) Including an entire chapter of this corpus within a publication to illustrate certain linguistic phenomena that you have annotated can also be problematic from the point of view of copyright. Though it is acceptable to quote portions of a text while indicating its source, these quotations should not exceed a clearly defined size limit. This size varies from country to country, but generally does not exceed a few hundred words. Thus, publishing an entire chapter of a book is not acceptable.

6.9. Further reading

Wynne (2005) addresses the different stages and difficulties associated with corpus data collection in an accessible but detailed manner. The methodological principles related to the creation of a corpus are discussed in detail by McEnery *et al.* (2005, section A) and more succinctly by Nelson (2010). Koester (2010) includes practical advice for the creation of a small specialized written corpus and Adolphs and Knight (2010) for the compilation of spoken corpora. The ethical and legal questions associated with the creation of a corpus are addressed by McEnery and Hardie (2012, Chapter 3).