1

Linguistic Resources for NLP

Today, the use of good linguistic resources for the development of NLP systems seems indispensable. These resources are essential for creating grammars, in the framework of symbolic approaches or to carry out the training of modules based on machine learning. However, collecting, transcribing, annotating and analyzing these resources is far from being trivial. This is why it seems sensible for us to approach these questions in an introduction to NLP. To find out more about the matter of linguistic data and corpus linguistics, a number of works and articles can be consulted, including [HAB 97, MEY 04, WIL 06a, WIL 06b] and [MEG 03].

1.1. The concept of a corpus

At this point, a definition of the term *corpus* is necessary, given that it is central for the subject of this section. It is important to note that research works related to both written and spoken language data is not limited to corpus linguistics. It is actually possible to use individual texts for various forms of literary, linguistic and stylistic analyses. In Latin, the word corpus means *body*, but when used as a source of data in linguistics, it can be interpreted as *a collection of texts*. To be more specific, we will quote scholarly definitions of the term *corpus* from the point of view of modern linguistics:

– A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language [CRY 91].

– A collection of naturally occurring language text, chosen to characterize a state or variety of a language [SIN 91].

– The corpus itself cannot be considered as a constituent of the language: it reflects the character of the artificial situation in which it has been produced and recorded [DUB 94].

From these definitions, it is clear that a corpus is a collection of data selected with a descriptive or applicative aim as its purpose. However, what exactly are these collections? What are their fundamental properties? It is generally thought that a corpus must possess a common set of fundamental properties, including representativeness, a finite size and existing in electronic format.

The problem with the representativeness of a corpus has been highlighted by Chomsky. According to him, certain entirely valid linguistic phenomena exist which might never be observed due to their rarity. Given the infinite nature of language due to the possibility of generating an infinite number of different sentences from a finite number of rules and the constant addition of neologisms in living languages, it is clear that whatever be the size of a corpus, it would be impossible to include all linguistically valid phenomena. In practice, researchers construct corpora whose size is geared to the individual needs of the research project. Thus, the phenomena that Chomsky is talking about are certainly linguistically valid from a theoretical point of view but are almost never used in everyday life. A sentence that is ten thousand words long and formed in accordance with the rules of the English language is of no interest to a researcher who is trying to construct a machine translation system from English to Arabic, for example. Furthermore, we often talk about applications which are task orientated, where we are looking to cover the linguistic forms used in an applied context, which is restricted to hotel reservations or asking for tourist information, for example. In this sort of application, even though it is impossible to be exhaustive, it is possible (even though it takes a lot of work) to reach a satisfactory level.

Often, the size of a corpus is limited to the given number of words (a million words, for example). The size of a corpus is generally predetermined in advance during the design phase. Sometimes, teams, such as Professor John Sinclair's team at the University of Birmingham in England, update

their corpus continuously (in this case, the term *text collection* is preferred). This continuous updating is necessary to guarantee the representativeness of a corpus across time: the opening up and the infinity of the corpus constitute a means to guarantee diachronic representativeness. Infinite corpora are particularly useful for lexicographers who are looking to include neologisms in new editions of their dictionaries.

Today, the word corpus is almost automatically associated with the word digital. Historically, the term referred mainly to printed texts or even manuscripts. The advantages of digitalization are undeniable. On the one hand, research has become much easier and results are obtained more quickly and, on the other hand, annotation can be done much more flexibly. Moreover, sometimes long-distance teamwork has become much easier. Furthermore, in view of the extreme popularity of digital technology, having data in an electronic format allows such data to be exchanged and allows paper usage to be reduced (which is a good thing given the impact of paper usage on the environment). However, this gave birth to some long-term issues related to electronic corpora such as portability. With the development of operating systems and text analysis software, it sometimes becomes difficult to access documents that were coded with old versions of software with a format that is obsolete. To get around this problem, researchers try to perpetuate their data using independent versions of platforms and of text processing software. XML markup language is one of the main languages used for the annotation of data. More specialized standards such as the EAGLES Corpus Encoding Standard and XCES are also available and are under continuous development to allow researchers to understand linguistic phenomena in a precise and reliable way.

In the field of NLP, the use of corpora is uncontested. Of course, there is a debate surrounding the place of corpora within the approach to build NLP systems, but to our knowledge, everyone is in agreement that linguistic data play a very important role in this process. Corpora are also very useful within linguistics itself, especially for those who wish to carry out a study on a specific linguistic phenomenon such as collocations, fixed expressions, as well as lexical ambiguities. Furthermore, corpora are used more and more in disciplines such as cognitive science or foreign language teaching [NES 05, GRI 06, ATW 08].

1.2. Corpus taxonomy

To establish a corpus taxonomy, many criteria can be used, such as the distinction between spoken corpora, written corpora, modern corpora, corpora of an ancient form of a language or a dialect, as well as the number of languages in a given corpus.

1.2.1. *Written versus spoken*

This kind of corpus is made up of a collection of written texts. Often, corpora such as these contain newspaper articles, webpages, blogs, literary or religious texts, etc. Another source of data from the Internet includes written dialogues between two people communicating on the Internet (such as in a chat) or between a person and a computer program designed specifically for this kind of activity. Often, newspaper archives such as *The Guardian* (for English), *Le Monde* (for French) and *Al-Hayat* (for Arabic) are also a very popular source for written texts. They are especially useful within the fields of information research and lexicography. More sophisticated corpora also exist, such as the British National Corpus (BNC), the Brown Corpus and the Susanne Corpus, which consists of 130,000 words of the Brown Corpus which have been analyzed syntactically. Written corpora can appear in many forms. These forms differ as much at the level of their structures and linguistic functions as at the level of their collection method.

– Verbal dictations: these are often texts read by office software users to gather digital texts in the form of data. Speakers vary in age range and it is necessary to record speakers of different genders to guarantee phonetic variation. Sometimes, geographical variations are also included, for example (in case of American English), New York English versus Midwest English.

– Spoken commands: this kind of corpus is made up of a collection of commands whose purpose is to control a machine such as a television or a robot. The structures of utterances used are often quite limited because short imperative sentences are naturally quite frequently used. Performance phenomena such as hesitation, self-correction or incompleteness are not very common.

– Human–machine dialogues: in this kind of corpus, we try to capture a spoken exchange or a written exchange between a human user and a

computer. The diversity of linguistic phenomena that we are able to observe is quite limited. The main gaps come from the fact that machines are far from being as good as humans. Therefore, humans adapt to the level of the machine by simplifying their utterances [LUZ 95].

– Human–human dialogues mediated by machines: here, we have an exchange (spoken or written) between two different human users. The mediator role of the machine could quite simply involve transmitting written sequences or sound waves (often with some extent of loss in sound quality). Machines could also be more directly involved, especially in the case of translation systems. An example of such situation could be a speaker "A" who is speaking in French and this person who tries to reserve a hotel room in Tokyo by speaking to a Japanese agent (speaker B) who does not speak French.

– Multimodal dialogues: whether they are between a human and a machine or mediated by a machine, these dialogues have the ability to combine gestures and words. For example, in a drawing task, the user could ask the machine to move a blue square from one place to another. Put this square <pointing gesture towards the blue square> here <pointing gesture towards the desired location>.

1.2.2. *The historical point of view*

The period that a linguistic corpus represents can be considered as a criterion for distinguishing between corpora. There are corpora representing linguistic usage at a specific period in the history of a given language. The data covered by ancient texts often consist of a collection of literary texts and official texts (political speeches, archives of a state). In view of the fleeting nature of oral speech, it is virtually impossible to accurately identify all the sensitivities of a spoken language long ago.

1.2.3. *The language of corpora*

A corpus must be expressed in one or several languages. This leads us to need to distinguish between: monolingual corpora, multilingual corpora or parallel corpora.

Monolingual corpora are corpora whose content is formulated with the help of a single language. The majority of corpora that are available today are of this type. Thus, examples of corpora of this type are very common: the Brown Corpus and the Switchboard Corpus for written and spoken English, respectively, and the Frantext corpus, as well as the OTG corpus for written and spoken French, respectively.

Furthermore, parallel corpora include a collection of texts where versions of the text in several languages are connected to one another. These corpora can be represented as a graph or even a matrix of two dimensions n x m: where *n* is the number of texts (*Tx*) in the source language and *m* is the number of languages. News reports from press agencies such as Agence France-Presse (AFP) or Reuters are classic examples of sources of such corpora: each report is translated into several languages. Furthermore, several organizations and international companies such as the United Nations, the Canadian Parliament and Caterpillar have parallel corpora for various purposes. Some research laboratories have also collected this type of corpora, such as the European corpus CRATER by the University of Lancaster, which is a parallel corpus in English, French and Spanish. For a corpus to really be useful, fine alignments must be made at levels such as sentence or word. Thus, each sentence from text "T1" in language "L1" must be connected to a sentence in text "T2" in language "L2". An extract from a parallel corpus with aligned sentences is shown in Figure 1.1.

sub d = 22 ----------&
the location register should as a minimum contain the following information about a mobile station :
-----&
1'enregistreur de localisation doit contenir au moins les renseignements suivants sur une station mobile:
sub d = 386 ----------&
handover is the action of switching a call in progress from one cell to another (or radio channels in the same cell).
-----&
le transfert intercellulaire consiste à commuter une communication en cours d'une cellule (ou d'une voie radioélectrique à l'autre à l'intérieur de la même cellule).

Figure 1.1. *Extract from a parallel corpus [MCE 96]*

Note that a multitude of multilingual corpora exist which are not parallel corpora. For example, the corpus CALLFRIEND Collection is a corpus of telephone conversations available in 12 languages and three dialects, and the corpus CALLHOME is made up of telephone conversations available in six languages. In these two corpora, the dialogues, which are not identical from one language to another, are not connected in the same way as in the format presented above.

Parallel corpora are a fundamental source used to build and test machine translation software (see [KOE 05]). An important question to ask after having identified multilingual data is the alignment of the content of these data. To resolve such a fundamental problem to make use of multilingual corpora, a number of approaches have been proposed. Some approaches are based on the comparison of the length of sentences in terms of the number of characters they contain [GAL 93] and in terms of the number of words [BRO 91], while others adopt the criterion of vectorial distance between the segments of the corpora considered [FUN 94]. Furthermore, there are approaches which make use of lexical information to establish links between two aligned texts [CHE 93]. Other approaches combine the length of sentences with lexical information [MEL 99, MOO 02]. Note that the GIZA++ toolbox is particularly popular for aligning multilingual corpora.

1.2.4. *Thematic representativity*

This criterion affects written corpora which target the representativity of an entire language or at least a large proportion of this language. To achieve representativity at such a broad level, having a selection of texts coming from a variety of domains is essential. Three types of layouts can be cited:

– Balanced corpora: to guarantee thematic representativeness, texts are collected according to their topics, so as to ensure that each topic is represented equally.

– Pyramidal corpora: in these cases, corpora are constructed using large collections for topics considered central and small collections for topics considered less important.

– Opportunistic corpora: this kind of corpora is used in cases where there are not enough linguistic resources for a given language or for a given application. Therefore, it is indispensable to make the most of all available resources, even if they are not sufficient to guarantee the representativeness aimed for.

Note that guaranteeing the topic representativity of a corpus is often complicated. In most cases, texts look at several different topics at once and it is difficult (especially in the case of an automatic collection from a corpus, with the help of a web crawler, for example) to decide exactly what topic a given text covers. Moreover, as [DEW 98] underlines, there is no commonly accepted typology used for the classification of texts. Finally, it may be useful to mention that lexicography and online information research are among the areas of application which are the most sensitive to thematic representativeness.

1.2.5. *Age range of speakers*

The application or scientific domains often impose constraints regarding the age range of speakers. Certain corpora are only made up of linguistic productions uttered by adult speakers, such as air travel information system (ATIS), distributed by LDC. Certain corpora that will be used to research first language acquisition are made up of baby utterances. The most well-known example of this is the child language data exchange systems (CHILDES) corpus, collected and distributed at Carnegie Mellon University in the United States. Finally, corpora exist which cover the linguistic productions of adolescents, such as the spoken conversation corpora collected at the University of Southern Demark SDU as part of the European project NICE.

1.3. Who collects and distributes corpora?

The increasingly central role of corpora in the process of creating AI applications has led to the emergence of numerous organizations and projects with a mission to create, transcribe, annotate and distribute corpora.

1.3.1. *The Gutenberg project*[1]

This is a multilingual library which distributes approximately 45,000 free books. This project makes an extensive choice of books available to Internet users, both at the linguistic level and at the level of topics available, since it distributes literary works, scientific works, historical works, etc. Nevertheless, since it is not specifically designed to be used as a corpus, the works distributed in this project need some preprocessing to make them usable as a corpus.

1.3.2. *The linguistic data consortium*

Founded in 1992 and based at the University of Pennsylvania in the United States, this research and development center is financed primarily by the National Science Foundation (NSF). Its main activities consist of collecting, distributing and annotating linguistic resources which correspond to the needs of research centers and American companies which work in the field of language technology. The linguistic data consortium (LDC) owns an extensive catalog of written and spoken corpora which covers a fairly large number of different languages.

1.3.3. *European language resource agency*

This is a European level centralized not-for-profit organization. Since its creation in 1995, the European language resource agency (ELRA[2]) has been collecting, distributing and validating spoken, written and terminological linguistic resources, as well as software tools. Although it is based in the European city of Paris, this organization does not only look at European languages. Indeed, many corpora of non-European languages, including Arabic, feature in its catalog. Among its scientific activities, the ELRA organizes a biannual conference: language resources and evaluation conference (LREC).

1 https://www.gutenberg.org/.
2 http://www.elra.info/en/.

1.3.4. *Open language archives community*

Open language archives community (OLAC[3]) is a consortium of institutions and individuals which is creating a virtual library of linguistic resources on a global scale and is developing a consensus on best practices for the digital archiving of linguistic resources by creating a network of storing services for these resources.

1.3.5. *Miscellaneous*

Given the considerable costs of a quality corpus and the lucrative character of most existing organizations, it is often difficult for researchers who do not have a sufficient budget to get hold of corpora that they need for their studies. Moreover, many manufacturers and research laboratories jealously keep back the linguistic resources they own, even after the projects for which the corpora were collected have finished.

To confront this problem of accessibility, many centers and laboratories have begun to adopt a logic that is similar to that of free software. Laboratories such as CLIPS-IMAG and Valoria have, for example, taken the initiative of collecting and distributing two corpora of oral dialogues for free. These corpora include the Grenoble Tourism Office corpus and the Massy School corpus[4] [ANT 02]. In the United States, there are examples such as the Trains Corpus collected by the University of Rochester, whose transcriptions have been made readily available to the community [HEE 95]. In addition, the *ngrams of the Google books*[5] is a corpus which is used more and more for various purposes.

1.4. The lifecycle of a corpus

As an artificial object, corpora can only very rarely exist in the natural world. Corpora collection often requires important resources. From this point of view, in some ways, the lifecycle of a corpus resembles the lifecycle of a piece of software. To get a closer look at the lifecycle of a corpus, let us examine the flowchart shown in Figure 1.2. As we can see that there are four

3 http://www.language-archives.org/.
4 http://www.info.univ-tours.fr/~antoine/parole_publique/Massy/index.html.
5 https://books.google.com/ngrams.

main steps involved in this process: preparation/planning, acquisition and preparation of the data, use of the data and evaluation of the data. It is a cyclical process and certain steps are repeated to deal with a lack of linguistic representativeness (often diachronic, geographical or empirical in nature) to improve the results of an NLP module.

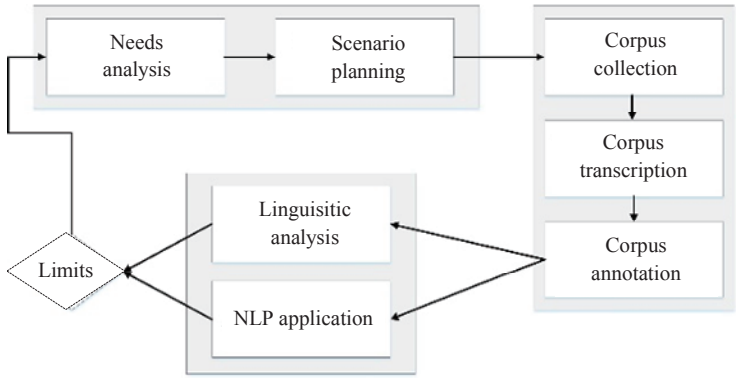

Figure 1.2. *Lifecycle of a corpus*

Three main steps stand out within a lifecycle:

– The preparatory step: this is about the work carried out before the corpus collection. In this step, key questions must be answered, such as: Why do we need a corpus? What properties should such a corpus have? How can we collect this corpus?

– The collection and the annotation of the corpus: this step covers the work necessary to construct the corpus in such a way that the objectives fixed in the preceding step can be reached.

– The use of the corpus: this step is about the statistical analysis and/or the linguistic analysis of the contents of the corpus. This step can bring some insights into the studied linguistic subject. For example, you can try to calculate the number of syntactic constructions by knowing the thematic context or the type of text (medical text, journalistic text, etc.). Moreover, the corpus can be used to construct NLP modules.

As we shall see later on, the lifecycle of a spoken corpus is distinguished by an additional step which is the transcription of spoken utterances. Moreover, given their situation in a specific spatio-temporal context,

dialogue corpora (both written and spoken) require the definition of scenarios to ensure a minimal level of representativeness of the dialogue domain.

1.4.1. *Needs analysis*

Examples of the objectives of a corpus include analyzing varieties of syntactic styles, constructing a morphological analyzer for a given language and creating a dictionary. The needs analysis directly affects all the parameters which define the type of a corpus. Among others, this allows the following to be decided:

– Basic choices: whether the corpus is spoken or written, the languages, etc.

– Speakers: the age range of speakers, their socioeconomic status, the number of speakers used, the gender of the speakers (percentage of males and females).

– Size of the corpus: when we have to collect a corpus to make a dictionary for the Arabic language, for example, we need to use a very broad corpus to make sure that all the linguistic registers and all socioeconomic factors have been taken into account.

– Thematic structure of the corpus: pyramidal, balanced, etc.

1.4.2. *Design of scenarios to collect data for the corpus*

After having specified the collection objectives, the linguists must describe how the corpus is to be collected. This must happen according to the objectives specified in the preceding step. Note that the scenarios used for collection involve both spoken and written conversation corpora and that one scenario can sometimes be adapted to several collection methods.

1.4.3. *Collection of the corpus*

As we have already seen, a corpus is a collection of texts that is specifically selected to satisfy a number of predetermined constraints. The simplest way of collecting a corpus is to use real existing data. As far as spoken data is concerned, the broadcast news is probably the most well-

known example. It consists of a televised news program accompanied by a written transcription. For written data, the Internet is incontestably the most abundant source. This is also reflected by the diversity of the linguistic forms and registers available online such as classical literature, informal chat and discussion forums.

Collection is carried out using a web crawler, which collects information automatically according to predefined thematic and linguistic criteria. Creating a list of documents can be done in two different ways. One way is to do this using a search engine: in this case, the crawler uses a number of keywords which it successively submits to one or several search engines. The URLs collected from the search results are added to the list of documents to be analyzed. The search engine plays the role of a topic filter here since only pages corresponding to the query topic are obtained. The other way is to obtain the list of documents using a list of URLs. This list can be initialized right at the beginning with a collection of links generated manually. Next, new URLs are extracted from the pages visited and are used to expand the list of URLs to be visited. This allows an exploration of the document space using a *Breadth First Search* approach. Note that crawlers must respect the rules of ethics which involve consuming the minimum amount of resources from the server from which the data are extracted. Often, crawlers are equipped with a language detection algorithm. An algorithm like this is able to classify the documents according to the language they are written with. Thus, the language and the theme of the text are, in general, the main selection criteria for a page to be included in a database. NLP specialists have made use of this source of information in the development of several types of applications, including speech recognition software and the POS tagging (see [VAU 00]).

In some cases, linguists use computer programs to generate sentences which correspond mainly to syntactic criteria. Among the most well-adapted tools is definite clause grammar (DCG), developed using the PROLOG language (logic programming). Due to the limitations of current automatic generation systems, it is often considered to be costly to constrain the syntactic grammars used for this kind of objective using semantic criteria. Thus, such corpora are of no interest to linguistic research. Often, they are used to train speech recognition modules (in particular, statistical language models). The main aim of this method is to obtain a number of syntactically acceptable texts with a minimal amount of time and effort.

To collect linguistic data that conform to specific criteria, it is possible to create a description of the system's task, which can then be used as a support for the generation of data. For example, at the University of Aleppo, in the framework of the construction of the prototype of our system AraTis (airline reservation system in Arabic), we carried out data collection of this type, since at the beginning of the project, no linguistic data of this type were freely available. The advantage of this method is that no special preparations are required. The only requirements to collect data of a reasonable quality and quantity include having a clear description of the system's task and getting a sufficient number of speakers. The number of speakers varies naturally from one application to another when collecting data of reasonable quality and quantity. Previous works have shown that the task as well as the physical context influence the linguistic behavior of speakers [LUZ 95]. This limits the possibilities of using such data for the rapid development of prototypes since the statistical representativeness of the phenomena is not guaranteed. The Wizard of Oz method is often used to address these shortcomings.

To develop a human–machine dialogue system of any type, we need to model several sources of knowledge at different levels. This includes linguistic and metalinguistic knowledge, which involve a considerable number of factors which directly influence how conversations progress. Besides, this knowledge includes information about the speaker, the speaker's way of speaking, the speaker's linguistic level (whether they are native or foreign). In addition, this knowledge includes information about the conversation topic, how certain operations are carried out and knowledge of the physical context, i.e. where the dialogue takes place (e.g. at a train station, at an airport or at the workplace, etc.).

To take into consideration all the knowledge that we have just outlined and to simulate the behavior of speakers when faced with a real system before its creation, researchers use the Wizard of Oz method. The idea of this method is to put the participant in a context which makes him think that he is interacting with an intelligent computer program, but in reality, he is interacting with a fellow human who is simulating the reactions of the machine. This is shown in the diagram outlined in Figure 1.3.

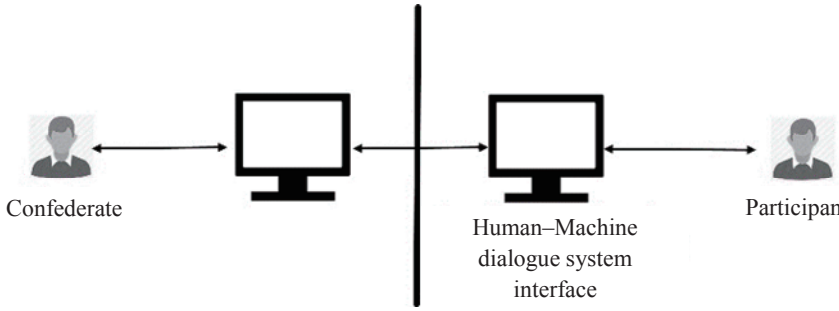

Figure 1.3. *Data collection system using the Wizard of Oz method*

The main advantage of the Wizard of Oz method is that it comes close to real utilization conditions and, therefore, the data produced is of a better quality, both linguistically and in terms of the knowledge linked to the applied usage of such data. However, in some cases, the cost and the tools necessary for collection can exceed the financial means of most laboratories. For some projects, such as those involving a dialog with an embedded system in a car or an airplane, we need to use simulators for these machines which makes the project extremely expensive. Therefore, only large specialized companies are able to carry out collection using tools of this kind (see [GEU 02] for an example of a collection for a corpus using a car simulator).

Manually collected corpora, or sometimes corpora collected using the Wizard of Oz method, are often used to develop a preliminary version of a system or a prototype. This prototype can be used to collect better quality data, which, in turn, can be used to improve the performance of the prototype itself (see Figure 1.4). For example, we can cite the Halpin system, which was developed within the laboratory CLIPS-IMAG [ROU 00]. This system of human–machine dialogue that can be used to research bibliographic references in the IMAG media library was put online to collect usage data. This data is used later to improve the quality of the system. Successive versions of the system were released, and at each iteration, the quality of the system improved and consequently the quality of the data collected was also improved.

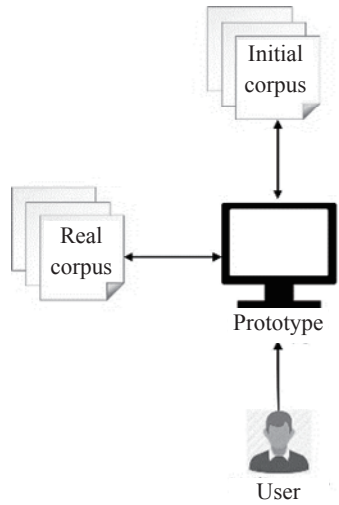

Figure 1.4. *Diagram of a corpus data collection system using a prototype*

Therefore, this is an incremental process that can continue as long as the system is in use. In this way, it is possible to take into account the potential evolutions of the linguistic and interactive behavior of users. Using prototypes for collection is a very good way of obtaining real data easily. On the other hand, this method requires a lot of resources to annotate the large quantities of data which are obtained in this way. Furthermore, since the prototype is made readily available to users, some users occasionally take the system for a game and, therefore, do not produce utterances that correspond to the purpose of the system. To filter out such utterances, extra effort is required.

1.4.4. *Transcription*

Transcription involves producing a written version of a recording obtained using one of the different collection methods. A professional transcription must be carried out rigorously and three fundamental principles must be respected [EDW 93]:

– categories must be discriminating, exhaustive and contrastive;

– transcriptions must be easy to read;

– transcriptions must be systematic and predictable to make the automatic processing of data possible.

Before beginning the transcription, the type of transcription must be decided, in order to know whether an orthographic, phonetic or prosodic (or a combination) transcription is required. If a combination is required, the transcriptions must be aligned. An agreement must be reached as much for the language in question as for foreign words regarding problematic spellings, which can be quite common in transcriptions, e.g. alternative spellings for Kuwait in the French language (*Kuweit*, *Koweït* and *Koweit)*. This is necessary to guarantee the homogeneity of transcriptions. In the same way, it is important to plan to take into account non-verbal phenomena present in the speech signal when they are produced by speakers, e.g. clicks, coughs, hesitations and long or short pauses. Short pauses are typically between 0.2 and 0.5 seconds while long pauses are those whose length exceeds 0.5 seconds. Equally, it is possible to consider sound phenomena linked to the environment where the conversation is being recorded such as objects falling, parallel conversations and the noise of cars or airplanes.

Often, the software used in transcription offers an open list in which the user can insert labels to be used for certain phenomena. Let us look at Figure 1.5, which gives us an example of the transcription of a radio sequence produced using the transcription software Transcriber[6], which is distributed under the general public license (GPL). In this example, each speaker's contribution begins with the name given to them by the transcribers. In the case of our example, we have two speakers: Simon Tivolle and Patricia Martin. The two first speaking turns are marked #1 and #2, respectively, which signifies that the two turns are happening in parallel. The labels [laugh] and [i] indicate, respectively, that a laugh and an inhalation occurred at that moment (by their presence in the sequence). Finally, the labels [laugh-] and [-laugh] show that the sequence between them is produced in parallel with a laugh.

Simon Tivolle : #1 yeah. #
Patricia Martin : #2 sure ? #
Simon Tivolle : really? [laugh] no. joke, Patricia's joke. [i] France-Inter, [laugh-] it's 7 o'clock [-laugh].
Patricia Martin : the news, Simon Tivolle:
Simon Tivolle : [i] hello! Tuesday April 28th. The national consultation on the national high school: [i] a huge debate today and tomorrow in Lyon to learn about

Figure 1.5. *Transcription example using the software Transcriber*

6 http://www.etca.fr/CTA/gip/Projets/Transcriber/.

Finally, note that the process of transcribing large amounts of data requires the implementation of a hierarchical cooperation process between several linguists to verify the transcriptions more than once and, therefore, ensure that the quality required is achieved.

1.4.5. *Corpus annotation*

Annotation is the process which involves enhancing the text with linguistic information or sometimes general information that describes the contents of the corpus. In other words, annotation involves adding value to the corpus, since it improves its quality and, therefore, opens up the ways in which the corpus can be used (see [PAL 10] and [PUS 12] for a general introduction to this). Annotation typically corresponds to the levels of linguistic structure: morphology, syntax, semantics, etc. The annotation of a corpus with non-linguistic information is also possible. Annotation can be carried out manually when appropriate, but very often, NLP tools are used to carry out annotation automatically. In this case, a checking and error correction phase is indispensable. A good annotation must always be well documented to guide users. It must be as neutral as possible regarding theoretical controversies to maximize the scope of its usage.

The first step in the annotation process is the raw corpus made up of *tokenized* but unannotated texts which are cleaned to remove special characters, if necessary. Sometimes, depending on the type of text, titles and paragraphs are marked.

Texts annotated with parts of speech are one of the most commonly used corpora. This kind of corpora are annotated using POS tags. This corpora is mainly used to build and test parts of speech taggers or to test syntactic parsers. An example of a fragment of text annotated using parts of speech is shown in Figure 1.6.

a. SpeakerB3/SYM./.
b. Well/UH what/WP do/VBP you/PRP think/VB about/IN the/DT idea/NN of/IN ,/, uh/UH ,/, kids/NNS having/VBG to/TO do/VB public/JJ service/NN work/NN for/IN a/DT year/NN?/.

Figure 1.6. *Segment of a corpus analyzed using parts of speech*

As we will see in the chapter of syntax, statistical parsing has made significant progress, especially in terms of robustness and the resolution of ambiguity, thanks to the availability of syntactically annotated corpora. In practice, the realization of these parsers requires syntactically parsed data which are commonly named treebanks. A grammar based parser is usually used to annotate the corpus syntactically. Next, linguists begin to review the annotated corpus to be able to correct the inevitable errors introduced by the parser.

To be widely usable, it is important these corpora are independent of existing syntactic theories. However, there are two main schools of thought within the linguistics community, namely the structuralism and the functionalism, mirroring the famous schism in syntax.

The structuralism focuses on noun phrases, verb phrases, etc. The *Penn Treebank* by the University of Pennsylvania is the most popular example of this type of treebanks [MAR 94]. It is made up of a syntactically annotated collection of sentences from the Brown Corpus and the Switchboard Corpus. An example of a simple sentence from the Penn Treebank is shown in Figure 1.7.

```
( (CODE SpeakerB1 .))
( (INTJ Okay . E_S))
( (CODE SpeakerA2 .))
( (INTJ Okay . E_S))
( (CODE SpeakerB3 .))
( (SBARQ (INTJ Well)
                (WHNP-1 what)
                (SQ do
                   (NP-SBJ you)
                   (VP think
                               (NP *T*-1)
                               (PP about
                                  (NP (NP the idea)
                                              (PP of
                                                 ,
                                                 (INTJ uh)
                                                 ,
                                                 (S-NOM (NP-SBJ-2 kids)
                                                             (VP having
                                                                         (S (NP-SBJ *-2)
                                                                           (VP to
                                                                              (VP do
                                                                                          (NP public
service work))))
                                                                         (PP-TMP for
                                                                                      (NP a
year)))))))))
```

Figure 1.7. *Extract from the Penn Treebank*

As we can see in Figure 1.7, the sentences are labeled in the style of the programming language Lisp rather than XML.

A tree corpus for French was also constructed at the Formal Linguistics Lab (LLF) at Denis-Diderot University in Paris [ABE 03]. Made up of about 22,000 sentences and 870,000 words, this corpus was created by extracting sections of the daily newspaper *Le Monde* that appeared in 1990, 1992 and 1993. The corpus covers texts written by a number of authors on varying subjects from economics to literature and politics, etc. In contrast to the Penn Treebank, this corpus used a format based on XML, as shown in Figure 1.8. It has been distributed freely since 2001.

```
<SENT nb="7">
<PP fct="MOD"> Parmi
        <NP> les candidats
        <PP>à
                <NP> la commission exécutive
                        <PP> de <NP> La CGT </NP>
                </PP>
        </NP>
        </PP>
</NP>
</PP> ,
<VN fct="SUJ"> on compte </VN>
<NP fct="OBJ"> quarante---quatre nouveaux---venus </NP>.</SENT>
```

Figure 1.8. *Extract from a tree corpus for French*

Functional annotation uses a radically different approach and focuses on syntactic relationships and dependencies between words. This is the case in the *Prague Dependency Treebank* and the *English Dependency Treebank* [HAJ 98]. In fact, [XIA 01] showed that it is not possible to convert a dependency tree corpus into a corpus annotated using the structural approach such as the Penn Treebank because the functional approach treats the subject and object equally regarding their attachment to the verb.

There are corpora which are semantically annotated. In contrast to syntactic annotation, semantic annotation approaches are quite diverse and fulfill a number of purposes. Some annotations cover semantic relationships between constituents in the sentence, e.g. the Proposition Bank [PAL 05].

Annotated at the University of Lancaster in the UK, the clinical text corpus, CLEF, is another example of a corpus of this type. Among the semantic relationships considered by this corpus, there is the *has_target* which compares an intervention or an investigation using the part of the corpus in question. It is, therefore, a predicate (relationship) which takes two arguments. The first argument is *investigation* or *intervention* and the second is *zone*.

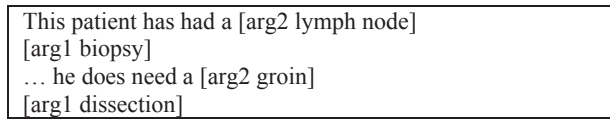

Figure 1.9. *Semantic annotation with a has_target relationship*

In the first sentence of the example shown in Figure 1.9, the predicate is *has had*, the intervention is *biopsy* and the zone of intervention is *lymph node*. The corpus GENIA is another semantically annotated medical corpus. It is a corpus which will be used to facilitate the extraction of knowledge based on genetic data [KIM 03]. Another form of annotation involves using temporal expressions such as those in the TimeBank [PUS 03].

There are also corpora which are annotated with discursive relationships, for example the RST Corpus, which is made up of 385 articles extracted from the Penn Treebank[7]. It is hierarchically annotated according to the rhetorical structure theory (RST) by [MAN 88]. The main task involved in annotation consists of identifying the elementary discursive units (EDUs). The discursive tree corpus Discourse Treebank from the University of Pennsylvania adopted an approach which was more centered on discursive connectors and their arguments [MIL 04]. It is probably useful to mention the annotation of co-referential relationships in the corpus by [POE 04] and the corpus of opinions [WIE 05].

Finally, it is probably worth mentioning some existing annotation tools. EXMARaLDA[8] is a German multi-level annotation tool which is entirely based on XML language. Specially adapted to discursive annotation, it contains a data annotation tool, a corpus manager which combines annotated files and adds the metadata. Developed at the Universidad

7 http://www.isi.edu/~marcu/discourse/Corpora.html.
8 http://www.exmaralda.org/.

Autónoma de Madrid, the UAM Corpus Tool[9] is another annotation tool, designed to be user-friendly to make annotation easier for linguists whose programming skills are limited [O'DO 08]. It is distributed with a number of NLP and research tools for English. The Brat Rapid Annotation Tool[10] by MIT is another example of an annotation tool. With a web interface, it is particularly adapted to collaborative annotation projects. It was used in projects about entity and event detection and extraction, as well as in projects about shallow parsing, etc. Other tools whose aims are more specific should also be mentioned. For example, CLaRK[11] for the annotation of syntactic information, NITE[12] for multimodal annotations and MMAX2[13] for anaphor annotation.

1.4.6. *Corpus documentation*

The aim of the documentation is to make corpora accessible to the community. Typically, three files are used to document corpora. Firstly, there is the initial file which is commonly called *readme*. This file contains information about the rights of authors, the version of the corpus, information about the corpus documentation (the other files) as well as summary information about the corpus: the size, the number of speakers, structure, etc. This is followed by the documentation file which includes a detailed description of all aspects of the corpus. Among other things, this includes the recruitment criteria for participants (e.g. age range, socioeconomic status, etc.), the annotation procedure, the format used, the software used, the recording and metadata. Finally, specific documents are put together to cover specific aspects of the corpus such as the history of the corpus, internal publications on the corpus in the form of technical reports, etc.

1.4.7. *Statistical analysis of data*

The statistical analysis of data involves looking at the frequency, the mean and the median of particular phenomena such as the frequency of a

9 http://www.wagsoft.com/CorpusTool/.
10 http://brat.nlplab.org.
11 http://www.bultreebank.org/clark/index.html.
12 http://www.ltg.ed.ac.uk/NITE/.
13 http://mmax2.net.

certain word or word category, a syntactic structure, an opinion, or another discursive phenomenon. It is possible to carry out the description of a given corpus or to compare these phenomena in two or several corpora.

1.4.8. *The use of corpora in NLP*

The way in which corpora are used to construct an NLP module depends on the approach used for processing. Rule-based approaches do not require specific annotations, since it is the responsibility of the human developer to extract the knowledge from the corpus as he or she sees fit. In contrast, learning-based approaches require annotated data to guide the process of information extraction and processing. The degree of granularity of the annotation required varies considerably according to the applicative aim of the module, as well as the algorithm and the approach that it adopts, such as whether it involves supervised or unsupervised learning, neural networks, statistical algorithms, and automatic grammar induction algorithms.

1.5. Examples of existing corpora

1.5.1. *American National Corpus*

This non-free corpus has the objective of collecting a million words from transcribed spoken data, as well as a collection of written texts whose size is approximately ten million words. The American National Corpus (ANC) team is made up of people in industry, as well as academic teams. This corpus includes important sections that are annotated with POS tags and is distributed using the XML coding standard format (XCES).

1.5.2. *Oxford English Corpus*

The Oxford English Corpus (OEC) is a collection of English texts which was used to support the creation of the *Oxford English Dictionary*, published by Oxford University Press. Containing more than two billion words, it is the largest corpus of its kind in the world. The texts which make up this corpus are extremely varied. Literary texts, specialized newspapers, daily newspapers, weekly newspapers, websites, and extracts of forums, among other types of texts, make up the main source of this corpus.

The OEC is annotated with XML and is often analyzed with the software Sketch Engine. Each document of the OEC is accompanied with the following metadata:

– title;

– author (if known);

– type of author (if known);

– dialect (British English, US English, etc.);

– source (website);

– date of the document (if known);

– date it was added to the corpus;

– field and sub-field;

– document statistics (number of tokens, sentences, etc.)

1.5.3. *The Grenoble Tourism Office Corpus*

Recorded by the laboratory CLIPS-IMAG in the Grenoble Tourism Office, this is a collection of task-oriented human–human spoken dialogues which come from the applied setting of tourist information [ANT 02]. The collection of data is carried out in real conditions following a semi-blind method: it involves an interaction between a member of the tourism office team and members of the public who are visiting the town. The real life conditions for recording meant that some sound quality was lost. The recordings were carried out on two different paths using a digital audio tape (DAT) recorder. In this way, two audio files in .wav format were obtained per conversation. In total, seven hours of recording were obtained. This corpus was initially limited to being distributed to members of the ARC. Today, it is distributed in two formats, the transcribed corpus can be downloaded directly from a web page associated with the project PAROLE PUBLIQUE[14] and the complete corpus (transcription and audio files), due to the size of the audio files, is distributed on CDs by post.

14 http://www-valoria.univ-ubs.fr/antoine/parole_publique.