D3.1 Framework for Data Collection

WP3: Data collection technologies for multiple corpus-based approaches

Hernani Costa (ESR3)

LEXYTRAD, University of Malaga, Spain
hercos@uma.es

Contents

1 Introduction

In the last decade, there has been a growing interest in bilingual and multilingual corpora. Particularly, in translation their benefits have been demonstrated by several authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Corpas Pastor and Seghiri, 2009). Their objectivity, reusability, multiplicity and applicability of uses, easy handling and quick access to large volume of data are just an example of their advantages. Thus, it is not surprising that the use of corpora has been considered an essential resource in several research domains such as translation, language learning, stylistics, sociolinguistics, terminology, language teaching, automatic and assisted translation, amongst others. Millions of users have created billions of webpages in which they express their vision and knowledge about the world. This linguistic and cultural content is considered a golden mine for those working in areas like Natural Language Processing (NLP), Information Retrieval (IR), Text Mining and Machine Translation (MT).

It is already a fact that the Internet can be seen as a large multilingual corpus due to its huge number of multilingual websites, in which different pages can contain the same written text in different languages. This means that some of their webpages can be paired into parallel texts, a very important source of knowledge for MT systems in general, and for Example-based Machine Translation (EBMT), Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT) in particular. Nevertheless, the lack of sufficient/up-to-date parallel corpora and linguistic resources for narrow domains and poorly-resourced languages is currently one of the major obstacles to further advancement on the aforementioned areas. One potential solution to the insufficient parallel translation data is the exploitation of non-parallel bilingual and multilingual text resources, also known as comparable corpora (i.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. EAGLES, 1996; Corpas Pastor, 2001:158). Even though comparable corpora can compensate for the shortage of linguistic resources and ultimately improve automated translations quality for under-resourced languages and narrow domains for example, the problem of data collection presupposes a significant technical challenge.

The solution proposed in iCorpora project and presented in this report is to exploit the fact that comparable corpora are much more widely available than parallel translation data. This ongoing project aims to increase the flexibility and robustness of the compilation, management and exploration of both comparable and parallel corpora by creating a new web-based application from scratch. iCorpora intends to fulfil not only translators' and interpreters' needs, but also professionals' and ordinary people's, either by breaking some of the usability problems found in the current compilation solutions available on the market or by improving their limitations and performance issues.

The remainder of the report is structured as follows. Section 2 introduces some fundamental concepts related with corpus linguistics, concepts of corpus linguistics and corpus compilation. Section 3 briefly presents a set of works that can be somehow related to this research. Then, section 4 describes a set of current compilation solutions available on the market, either to compile comparable corpora or to mine parallel texts from the Web. Finally, before presenting the final remarks and highlight some ideas in section 6, the necessary steps to design and develop a robust and agile web-based application to semi-automatically compile, manage and explore both parallel and multilingual comparable corpora is presented in section 5.

2 Background

The interest in mono-, bi- and multilingual corpora is vital in many research areas such as language learning, stylistics, sociolinguistics, translation studies, amongst other research areas. This section aims to present the fundamental concepts about Corpus Linguistics, their advantages and drawbacks, as well as their importance in various research areas (section 2.1, 2.1.1 and 2.1.2, respectively).

2.1 Corpus Linguistics

Corpus linguistics can be defined as the study of language and a method of linguistic analysis which uses a collection of "real world" texts called corpus (cf. McEnery et al., 2006; Taylor, 2008; Lüdeling and Kytö, 2008). It also can be seen as "*a tool, a method, a methodology, a methodological approach, a discipline, a theory, a theorical approach, a paradigm (theoretical or methodological), or a combination of these*", as pointed out by Taylor, 2008. Corpus linguistics aims to analyse and investigate various linguistic questions, such as how language varies from place to place, determine how specific words and their synonyms collocate and vary in practical use, amongst other questions that will be later addressed in detail. Due to the fact that this study offers an unique view to the language dynamism, it is not surprising that corpus linguistics is considered one of the most widely used methodologies since the early 20^{th} century (cf. Firth, 1935) and, one of the fastest-growing methodologies in contemporary linguistics (cf. Gries, 2009:1). Its history can be divided into two periods, the early corpus linguistics, also known as *pre-Chomsky corpus linguistics*, and the *modern corpus linguistics* (before and after the middle of 1980s, respectively).

Early studies based on corpus linguistics date to the first half of the 20^{th} century, where linguistics like Franz Boas, Leonard Bloomfield, John Rupert Firth and Zellig Sabbettai Harris (Boas, 1911; Bloomfield, 1933; Firth, 1935; Harris, 1951; Fries, 1952; Firth, 1957a;b), amongst many others, used corpus-based approaches to study the language. The first researchers in corpus linguistics defended the idea that the meaning of a word depended on their co-occurrence with other words and consequently it would lead to their contextual concept of its lexical meaning, as stated by Firth, 1957a:11: "*you shall know a word for the company it keeps*" or other authors like Cruse, 1986 and Wanner, 1996. Nevertheless, the arrival of the structuralists in the late 1950s, being Noam Chomsky one of the most influential in that period (cf. Chomsky, 1955; 1957; 1965; 1975; 1986; 1993), represented a change in orientation toward rationalism, which would result in a period of decline to corpus linguistics as an empiricist method. Therefore, in the following two decades (i.e. during the 1960s and 1970s), corpus linguistics was heavily criticised, especially from the practical point of view of rationalists like David Abercrombie (Abercrombie, 1965) – mostly due to the non-existent of required elements to process the data. At that time linguists worked with corpora manually, which, as we can imagine it was a tedious and dubious reliable approach. Thus, Chomsky rejected the use of corpus as a tool for linguistic studies, arguing that the linguist must model language on competence instead of performance (Chomsky, 1957; 1965; 1986; 1993). According to him, corpus does not allow language modelling on competence: "*Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.*" (Chomsky, 1962:159).

Even though corpus linguistics was not completely abandoned, it was not until the 1980s that linguists began to show an increased interest in the use of corpus for research. The revival of corpus linguistics and its emergence in the modern form was mostly influenced by the advent of computers and network technology in the aforementioned decade. These two powerful instruments gave linguists, for the first time in history, the necessary electronic tools to "easily" create, storage, and handle large collections of electronic data. As a result of these advances, several corpora, previously manually created were converted to electronic format – see

for instance the: Survey of English Usage Corpus (SEU)[1], compiled in 1959 by Quirk (University College, London); Brown University Standard Corpus of Present-Day American English (or just Brown Corpus), compiled in 1961 by W. Nelson Francis and Henry Kučera at Brown University, Providence, Rhode Island (Francis and Kučera, 1979); and the Lancaster-Oslo/Bergen Corpus (LOB), also compiled in 1961 by Stig Johansson and Geoffrey N. Leech in collaboration between the University of Lancaster and the University of Oslo (Johansson et al., 1978; 1986).

All the works mentioned before the 1980s, as well as the early examples of corpus linguistics, paved the way to modern study of language based on corpora as we know it today. In fact, it was only in 1984 that Aarts and Meijs coined the term "*corpus linguistics*", in the work entitled "*Corpus linguistics: Recent developments in the use of computer corpora in English language research*" (Aarts and Meijs, 1984). After that, several authors attempted to give a rigorous definition to the term, as being: "*the study of language on the basis of text corpora*" (Aarts, 1991:1) or "*the use of large collections of text available in machine-readable form*" (Svartvik, 1992:7). Despite of the no existence of a consensus about the definition, at that time a corpus could be characterised as a large collection of electronic data, publicly available and, traditionally known as *monitor corpus* because it had a closed set of textual material. Additionally, when more than one language are brought together, for instance to enable research on translation, language teaching, etc., it was called *multilingual corpus*. As we are shall see, more rigorous definitions will be written between the end of the 20^{th} century and the beginning of the 21^{st} century. It is interesting to mention that the advances in hardware and software at that time brought new programs capable of automatically recognise named entities and disambiguate, for instance, word-categories and consequently new corpora were brought to light, such as The Bank of English[2] (also known as COBUILD) in 1987, the International Corpus of English (ICE)[3] in 1990, and The British National Corpus (BNC)[4] in 1991, amongst others corpora.

Another important fact that caused a great interest in mono-, bi- and multilingual corpora was the demand by our society for texts in several languages. An interest that started in the beginning of the 1990s and increased throughout the years up to the present day, currently being vital in many research and practical areas, such as language learning, sociolinguistics, stylistics, and particularly in the translation field. A fact that can be easily explained by the today's economic relations, which require the edition of documents in more than one language – imagine the amount of international organisations that have a need to publish documents in several languages, e.g. technical manuals, reports, books, journals, etc. This demand, coupled with the globalisation phenomenon, has led to an increasing interest in bi- and multilingual corpora by researchers in a variety of research fields, such as machine translation (Brown et al., 1993), language teaching (Botley et al., 1996; Wichmann et al., 1997), specialised languages (Thomas and Short, 1996), terminology (Wright and Budin, 1997), teaching and practice of specialised translation (Corpas Pastor, 2001; 2002). Thus, corpus linguistics has contributed to the advances in disciplines so diverse as natural language processing, language engineering, terminology, translation, amongst others as pointed out by Pérez Hernández, 2002:128-129: "*The research based on corpora has led to the emergence of new methods of study in a wide range of areas of study, so diverse as lexical knowledge extraction, construction of grammars, cultural studies, stylistics, machine translation, speech recognition, information retrieval, mono- and bilingual lexicography, electronic dictionaries construction or the compilation of computational lexicons and of course, the creation of terminological information repositories.*". In the same line of thought, Corpas Pastor, 2003 states that translation studies based on corpora have helped researchers and translators to reflect about the notion of equivalence, linguistics comparison and the nature and characterisation of the translationese, as well as about its application in other research fields. Moreover, considering the practical point of view of corpora in translation

[1] `http://www.ucl.ac.uk/english-usage`
[2] `http://www.titania.bham.ac.uk`
[3] `http://ice-corpora.net/ice`
[4] `http://www.natcorp.ox.ac.uk`

studies, Zanettin, 2002 states that: "*In the last few years information technology has brought about a completely new scenario. The availability of vast quantities of texts in many languages and on all kinds of subjects is a dream come true for translators as well as for all types of discourse professional, text processors and language services providers.*".

As we can see, researchers and teachers seem to agree on the importance of corpora in general, and virtual corpora (or *ad hoc*) in particular, within translation training and practice (cf. Laviosa, 1998; Corpas Pastor, 2001; Bowker and Pearson, 2002; Zanettin et al., 2003). But, it is important to mention that corpora only was applied in translation studies in 1993. It that year Mona Baker, for the first time, discusses how to apply corpus linguistics techniques and methods in order to study the nature of translated texts (cf. Baker, 1993). According to her, corpora can help theorists of translation to observe and explore translations: "*The profound effect that corpora will have on translation studies, in my view, will be a consequence of their enabling us to identify features of translated texts which will help us understand what translation is and how it works.*"(Baker, 1993:242-243). She also predicted that the availability of large corpora of both original and translated text, together with the development of a corpus-driven methodology, will enable scholars to uncover "*the nature of translated texts as a mediated communicative event*" (Baker, 1993:243). Despite corpus linguistics could allow automatic techniques to analyse large collections of texts, she argues that it should be carefully designed and compiled for a specific research goal. She particularly suggests the use of monolingual comparable corpora, describing it as a collection of texts in the same language divided in two categories: one comprising texts translated into that language, and the other comprising the "*original texts in the language in question*" (Baker, 1995:234). The availability of these techniques represented an opportunity for the advancement of research within the domain: "*With the availability of corpus techniques, we can now go a step further and look not just at the functional types of translation but at the distinctive features of translated text per se.*" (Baker, 1996:176). These findings had a positive and fruitful implications amongst research community (cf. Laviosa, 2004), as they opened up new research lines and methodologies (Baker, 1995) that could be used to "*study translation as a variety of language behaviour that merits attention in its own right*" (Baker, 1996:176).

Thus, since the beginning of the 1990s, corpus-based approaches have been widely "accepted" by the scholars and consequently became one of the most important methodologies in several areas, particularly for contemporary translation studies - it was even suggested that it was the "*major methodological advance associated with corpus studies*" (Pym, 2008:321-322). The most important advantage of this methodology is that does not require prior knowledge or familiarity with the target language and culture (McEnery et al., 2006:6). Nevertheless, a number of arguments against the corpus-based approach were also pointed out. For instance, Tymoczo, 1998 rejects this type of approach as a suitable mode of research because of the subjective judgement of the researchers at every stage, starting from the corpus compiling process to the result interpretation stage (Tymoczo, 1998). Olohan, 2004 and Chesterman, 2004 argue that it is unclear to ascertain to what extent some universals are due to contrasting linguistic systems or to the process. Yet, the reality is that the today's research on translational hypotheses and translationese are broadly based on the use of corpora, both parallel and comparable corpora. By a way of example, parallel corpora are used to investigate the process of translation, to analyse how a message is transmitted in the target language, whilst the comparable corpora approach is usually suggested in product-orientated investigations of translation.

To conclude, over the last few years the technological advancements and the greater collaboration between translation scholars and Information Technology (IT) experts have made possible to compile larger samples of data and store them in electronic form to create corpora, many of them publicly available to the community, which has certainly increased the quality of both translational studies and professional translations.

2.1.1 Advantages and Disadvantages

The advantages and disadvantages of using corpora have been argued in the literature by several authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Gries, 2009, amongst others). Hereafter, some of them are pointed out.

Apart from the main positive aspects previously mentioned (i.e. its objectivity, reuse, multiplicity and applicability of uses, easy handling and quickly access to large volume of data), corpus linguistics also:

- Empowers the study of the foreign language: the study of the foreign language with the use of corpora allows the foreign language learners to get a better "feeling" about that language and learn the language through "real world" texts rather than "controlled" texts (cf. Gries, 2008).
- Simplifies the study of naturalistic linguistic information: as previously mentioned, a corpus assembles "real world" text, mostly a product of real life situations, which results in a valuable research source for dialectology (cf. Hollmann and Siewierska, 2006), sociolinguistics (cf. Baker, 2010) and stylistics (cf. Wynne, 2006), for example.
- Helps linguistic research: as the time needed to find particular words or phrases has been dramatically reduced with the use of electronically readable corpora, a procedure that would take days or even weeks to be manually performed can be done in a couple of seconds or even milliseconds with an high degree of accuracy.
- Enables the study of wider patterns and collocation of words: before the advent of computers, corpus linguistics was studying only single words and their frequency. More recently, the emergence of modern technology allowed the study of wider patterns and collocation of words (cf. Roland et al., 2007).
- Allows simultaneous analysis of multiple parameters: in the last decades, the development of corpora linguistic software tools helped the researchers to analyse a wider number of parameters simultaneously, such as determine how the usage of a particular word and its syntactic function varies.

However, not everything can be advantages, and hereafter some of the main disadvantages of their usage are highlighted:

- Manual intervention is often required: sometimes it is necessary to resort to manual intervention, which adds some issues. The lack of an agreement on the size that a corpus should have or even its representativeness to the purpose for which it was compiled are just some examples of the difficulties involved in the process.
- Corpus linguistic studies do not explain the "why": the study of corpora answers "the what" and "the how" it happened, but it does not has the answers to "the why". For instance, corpus linguistics can not explain why the frequency of a particular word has increased or decreased over time.
- Corpora do not represent the entire language: corpus linguistics studies the language by using random or selected corpora, which typically assembles a large number of "real world" texts. However, these corpora do not represent the entire language.

2.1.2 Applications of Corpus Linguistics

Despite their disadvantages, our information society requires for texts in several languages, which together with the economic globalisation has brought a great interest in mono-, bi- and multilingual corpora. Indeed, nowadays, the use of corpora is vital in many research areas, like language learning, stylistics, sociolinguistics, translation studies, amongst other areas.

- Lexicography: in lexicography, corpus linguistics plays an important role in compiling, writing and revising dictionaries. Moreover, allows the linguist to get examples of words or phrases from millions of written texts in a few milliseconds. Additionally, due to the fact that corpora are constantly being updated or even expanded with new texts, lexicographers have a constant access to up-to-date information. For instance, in Summers, 2005 it is described how the frequency of words in various corpora has influenced the semantic description given in the definition, and the ordering of definitions in several entries in dictionaries.

- Grammar: likewise lexicography, grammatical studies rely heavily on corpus linguistics and the use of corpora. Even though corpora do not represent the entire language, the large volume of data offers a reliable representation of the language to conduct grammatical research, create/maintain grammars and also test theoretical hypotheses (cf. Aarts, 1991; Oostdijk and Haan, 1994; Hunston and Francis, 2000, amongst others).

- Sociolinguistics: as corpora assemble "real world" texts and often include all sorts of written language, ranging from literary works to the everyday language, they offer a valuable insight into how language varies from place to place and between different social groups. For example, Kjellmer (Kjellmer, 1986) used the Brown (Francis and Kučera, 1979) and LOB corpus (Johansson et al., 1978; 1986) to examine the masculine bias in American and British English. An overview on the ways that corpus linguistics approaches can be used in order to aid sociolinguistic research can be found in Baker, 2010.

- Translation studies: since corpora contain texts in different languages (i.e. bi- or multilingual parallel/comparable corpora), they are a valuable tool for translators as they can easily determine how specific words and their synonyms collocate and vary in practical use. The first author to discuss how to apply corpus evidence in order to study the nature of translated texts was Baker in 1993 (Baker, 1993). After her, a wide number of studies based on corpora and their effects on translation have influenced, not only the area of translation universals, but also other areas such as machine translation and translator training. Consequently, it had a significant impact on translation profession (cf. Baker, 1996; Bowker, 2002; Corpas Pastor and Seghiri, 2007b; Partington, 2011; amongst others).

- Language learning/teaching: a wide number of textbooks, which are used for language learning/teaching, contain texts from real contexts rather than being constructed for pedagogical purposes, giving the learners access to the facts of authentic language use. For instance, one of the most popular corpus used in language learning is the Spoken English Corpus (Taylor and Knowles, 1988). Moreover, corpus-based approaches have also been used on first and second language acquisition. For a general overview on first and second language acquisition using corpora, see Behrens, 2008 and Gries, 2008, respectively.

- Stylistics: it does not use corpora as often as other research areas because stylistics linguists are mostly interested in particular texts and authors. Nevertheless, corpora is seen as an important source of information by specialists in stylistics who are interested in wider genres, such as the language used by politicians, advertising industry, etc. For more information about corpus stylistics methods see Wynne, 2006.

- Dialectology: corpora have been an important source of research for dialectology for a long time. The "original form" of the texts included in corpora, including dialects, gives the linguists an invaluable perspective of geographical variation of a language (cf. Hollmann and Siewierska, 2006). Two examples of dialect corpora existing at present are the Helsinki corpus of English dialects (Rissanen et al., 1991; Kytö, 1996; Ihalainen et al., 2006) and Kirk's Northern Ireland Transcribed Corpus of Speech (Kirk, 1992).

- Historical linguistics: historical corpora assembles texts from specific historical periods or even "dead" languages, resulting in a *closed corpus* of data which is only extended by the (re-)discovery of previously unknown manuscripts or books. Historical corpora offer an easy access to these historic books and manuscripts in electronic form, which historical linguists consider a valuable source to their research. The most widely known English historical corpus is the Helsinki corpus (Rissanen et al., 1991; Kytö, 1996).

2.2 Concepts of Corpus Linguistics

This section presents a systematic methodology for automatic corpora compilation. Firstly, section 2.2.1 defines what a corpus is. Then, the corpus design criteria are described in section 2.2.2.

2.2.1 Definition of Corpus

Even though the term corpus has been used as a general term to define any compilation of text, a collection of texts is not *per se* a corpus. To be considered a corpus in the strict sense of the term, a set of clear design criteria must be stabilised and a systematic compilation protocol carried out (EAGLES, 1994; 1996a;b; Corpas Pastor, 2001).

Formalise the concept of corpus is not an easy task, nevertheless the definition proposed by John Sinclair in EAGLES, 1996b:4 is the most accepted in the research community: *"A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language."*. The author also defines the minimum criteria to be met by collections of texts, in electronic format, so these collections can be considered a proper corpus, namely: the quantity (the corpus size in number of words), quality (representativeness and balance), encoding simplicity and documentation (Ibid).

Nevertheless, a corpus should not be confused with other electronic collections (Atkins et al., 1992; Torruella and Llisterri, 1999:51-52), such as the *archive/collection* or the *electronic text library*.

- Archive/Collection: is a repository of readable electronic texts, not linked in any coordinated way, i.e. does not have any structure or linguistic criteria because the most important factor to its creation is the availability of the data.

- Electronic text library: is a collection of electronic texts in a standardised format with certain conventions related to the content, but without rigorous selectional constraints.

- Corpus: is a compilation of texts, but different to the previous electronic collections attends to specific linguistic criteria. It is codified following a standard and homogeneous process, allowing the study of the behaviour of one or more languages. In other words: "*Computer corpus: a corpus which is encoded in a standardised and homogeneous way for open-ended retrieval tasks.*" (EAGLES, 1996b:5).

It is also important to mention that a corpus can be divided into two more levels: *subcorpus* and *component* (cf. EAGLES, 1996a:5; Torruella and Llisterri, 1999:52):

- Corpus: as previously mentioned, a corpus can be seen as a set of pieces of language, selected and ordered according to explicit linguistic criteria with the purpose of represent a language or some part of a it (cf. EAGLES, 1996b:4).

- Subcorpus: a subcorpus inherits all the properties from the corpus as it is a part of a larger corpus.

- Component: a component is not necessarily an adequate sample of a language. It is a collection of pieces of language that are selected and ordered according to a set of

criteria with the purpose of characterise its linguistic homogeneity[5]. Whereas a corpus may illustrate heterogeneity[6], as well as a subcorpus to some extent, the component only illustrates a particular type of language.

Finally, relying on the variety of languages that can be identified in a corpus, the corpus can be called monolingual or multilingual corpus. A monolingual corpus is one that contains texts in a single language, while multilingual corpora contain texts in two or more languages. Specifically, a corpus built from collections of documents in two languages is called bilingual corpora, and when more than two languages are presented the corpus is called multilingual corpora.

2.2.2 Corpus Design/ Classification

As corpora have become larger, more diverse and widely used (e.g. they are more frequently used to make/take conclusions about the language), the used procedure to their compilation has become an important issue. Despite the absence of a well-defined design/classification criteria, one of the most complete proposals in the literature is the one proposed by Corpas Pastor, 2001 (see also Corpas Pastor, 2008; Corpas Pastor and Seghiri, 2009; Seghiri, 2011). In this work, the author combines different ideas proposed by several experts in the field (cf. EAGLES, 1994; 1996a; Baker, 1993; Johansson and Oksefjell, 1998; Torruella and Llisterri, 1999), in which the EAGLES reports were used as a starting point (EAGLES, 1994; 1996a), extended with the typology proposed by Torruella and Llisterri, 1999, and merged with the ideas of Baker, 1993 and Johansson and Oksefjell, 1998 about multilingual corpus classification. In the following topics five aspects of corpus design/classification are explained in detail: size, specificity, sample size, encoding and documentation.

Corpus Size The first classification criterion is related to the percentage and distribution of the different types of text contained in the corpus.

- Large Corpus: its size in not limited and it is usually composed by a large number of words. Another particularity of this type of corpus is their lake of representativeness and its unbalance sample sizes.
- Balanced Corpus: integrates several language varieties, in similar percentages.
- Pyramidal Corpus: the texts assembled are distributed by levels. These levels are characterised by the progressive increasing complexity of the topics included. In other words, the more complex the text is, the higher its level in the pyramid and more reduced the number of texts will be.
- Monitor Corpus: the number of texts in this corpus is invariable, but constantly updated, i.e. old texts are replaced by new texts whenever possible. Thus, this copus can be seen as a reference for the language evolution. Bowker and Pearson, 2002:12-13 named this corpus as "*open corpus*" due to its dynamism, and point out that: "*Given the dynamic nature of Language for Special Purposes (LSP) and the importance of staying abreast of current developments in the subject field, open corpora are likely to be of more interest for LSP users*".
- Parallel Corpus: is composed by collections of texts in one original language and its translations to one (or more) target language(s). When only two languages are involved, i.e. when the corpus has the original texts and its translation to a single target language, it is named bilingual parallel corpus. When more than two target languages are involved it is named multilingual parallel corpus. The most well known example is the Europarl Corpus (cf. Koehn, 2005).

[5]Homogeneity: the quality of being similar or comparable in kind or nature.
[6]Heterogeneity: the quality of being diverse and not comparable in kind.

- Comparable Corpus: is a corpus that includes similar types of original texts. As it is compiled from a original language in accordance with the same design criteria, these texts allow the comparison of their interlingual components (Corpas Pastor, 2001:158). Similarly to the parallel corpus, when only two languages are involved the corpus is named bilingual comparable corpora and multilingual comparable corpora when more than two languages are involved. In addition to these two subtypes, a third one named monolingual comparable corpora was been proposed by Corpas Pastor, 2001:158. Different from the first two subtypes this specific corpus includes original texts and their translated texts in the same language.

Corpus Specificity The second classification criterion classifies the corpus based on the included text specificity.

- General Corpus: as described by Bowker and Pearson, 2002:11-12, a general corpus is a corpus that "*can be taken as representative of a given language as a whole and can therefore be used to make general observations about that particular language*". It assembles, for example daily news or articles from newspapers, as its focus is the "*language for general purpose (i.e. the language used by ordinary people in everyday situations)*" (Bowker and Pearson, 2002:12). Nevertheless, Corpas Pastor, 2001:156 clarifies that besides general corpus there are also restricted corpus, such as specialised, generic, canonical, chronological and historical corpus. The author also pointed out that a general corpus should not be confused with lower levels of corpus as the subcorpus and the component.

- Specialised Corpus: is a corpus that is focused on a particular aspect of a language (Bowker and Pearson, 2002). Using the words from Bowker and Pearson, 2002:12: "*It could be restricted to the Language for Special Purposes (LSP) of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of a certain demographic group (e.g. teenagers).*".

- Generic Corpus: is a corpus that assembles samples from a particular gender.

- Canonical Corpus: is a corpus that contains complete works of an author.

- Chronological Corpus: is a corpus that contains texts that have occurred over a period of time. This type of corpus can be also referred as "*synchronic corpus*" (Bowker and Pearson, 2002:12).

- Historic Corpus: a corpus that includes texts from different periods of time with the purpose of carry out studies about the language evolution (Abaitua, 2002).

Corpus Samples Size The quantity of text used in the samples to assembly a corpus is the third classification criterion.

- Textual Corpus: is a corpus created by "whole text", with the purpose of represent the language, as well as their most important varieties. This type of corpus is broad used in the creation of grammars and dictionaries, for example.

- Reference Corpus: whereas textual corpus assembles "whole texts", a reference corpus is composed by samples of the "whole text". The aim is not in the text itself, but rather seeks to represent some particularity of a language or language characteristic.

- Lexical Corpus: is composed by small samples, with similar length, with a specific purpose, the lexical study.

Corpus Encoding The fourth classification criterion is related to the corpus encoding.

- Annotated Corpus: comprises information generated and added to the primary data as a result of some linguistic analysis (cf. tagsets for encoding linguistic annotation, that could be information about: segmentation of the text into sentences and words, morphosyntactic tagging, parallel text alignment, amongst other features).

  A Corpus Encoding Standard (CES)[7] has been developed to serve as a widely accepted set of encoding standards for corpus-based works. This standards specify a minimal encoding level that a corpus must achieve to be considered standardised. It also provides encoding specifications for linguistic annotation, together with a data architecture for linguistic corpora.

- Unannotated Corpus: most often created for non-linguistic purposes, such as publishing. This raw text corpus presents a high level of simplicity since has not been added any type of linguistic annotation. The most common format is plain text with the character-encoding scheme ASCII.

Documentation The fifth classification criterion is related with the corpus documentation.

- Corpus with documentation: to make best use of a corpus, it is necessary, not only have access to the texts, but also to the explanatory documentation, licence agreements, meta-data[8], etc., also known as corpus manifest. As far as possible, all such supporting documentation should be included along with the corpus itself. Usually the structure of a document is divided into two elements: the header that contains the meta-data and the body with the document content. For instance, the header could contain the following fields: title (the title of the document), author (the author of the document), year (publishing year), availability (either free or license), amongst others elements that help to describe the document structure. The document body contains text-entities and can also have sections. The basic text-entities could be lists, tables, paragraphs or other unformatted text. The sections have the purpose of separate the text-entities. There is a consortium named Text Encoding Initiative (TEI)[9] which purpose is the development and maintenance of a standard for the representation of texts in digital form.

- Corpus without documentation: as its name suggests, this type of corpus does not have any documentation associated.

2.3 Corpus Compilation

This section starts with section 2.3.1 describing in detail the compilation workflow, i.e. all the required steps to compile corpora from the Internet. Then, section 2.3.2 presents some ideas to access the degree of comparability that documents in corpora should have.

2.3.1 Protocol

After establishing the design criteria, the next stage is the compilation protocol, which can be divided into four steps, as proposed by Corpas Pastor, 2008 (see also Corpas Pastor and Seghiri, 2009 and Seghiri, 2011): finding data, downloading the data, normalisation and storage. An additional step should be considered in order to ensure the corpus representativeness, to the object of study.

[7] `http://www.cs.vassar.edu/CES`
[8] Meta-data: data that describes other data.
[9] `http://www.tei-c.org/index.xml`

Finding Data After establishing the design criteria, the next stage is the identification of reliable sources. There are basically two types of searches that can be made over the Internet to find these sources: institutional and thematic.

The institutional search is directed to institutional companies, organisations and institutions. The information available through this specialised web sources result in a high standard of quality and reliability as the writers are professionals and specialists in the field.

The thematic search is normally carried out by the use of search engines. Firstly a set of keywords is defined. Then, these keywords are combined along with truncations and boolean operators with the purpose of create search queries. At this point, it is very important to create well-define queries in order to avoid a large amount of irrelevant documents to be returned. Finally, the documents returned by these queries are manually analysed, and the irrelevant data is filtered out.

Downloading Data Once defined the target sources, the next stage is to retrieve the data from these sources. This process can be manually made or automatically by using computational programs. Hereafter a short explanation about the main approaches used to acquire data is presented.

- Existing Collections: this approach takes advantage of existing collections, handcrafted or automatically compiled. If by on one hand these collections provide an instant availability of linguistic data, on the other hand they are limited to its design constrains, resulting in a static and obsolete resource to specific demands.

- Web-based Approach: this approach was designed to overcome the problems in the previous approach, by taking advantage of all the resources available in the Internet. Nevertheless, it has several advantages and disadvantages. Some of the advantages are the availability of: massive amounts of electronic text, public domain documents, and wide reach of text-types/topics/genres/domains. The disadvantages are: the difficulty of copyright ascertainment (something that also occurs with the previous approach); additional effort to clean the documents' meta-data; the difficulty to achieve a balanced corpus; and finally, despite of the quantity of information at our disposal, the difficulty in selectively retrieve quality documents, increases. Despite the drawbacks this approach is widely used, not only by researchers but also by professional on their daily tasks. Usually one of the two Web-based approaches is used: Web Search Engine or Web Focused Crawling.

  - Web Search Engine: the aim of this approach is to search the Web for pages that contain information about a pre-defined topic (yet, it can be used to exploit corpus for any topic or domain). To do that, a well-defined set of keywords that characterise a specific topic/domain should be defined. Then, these keywords are converted into search query strings. With the purpose of create more accurate search queries, the keywords are combined with boolean operators in order to define a relationships between them. In order to harvest the resulting documents, these search query strings are submitted to a search engine. The quantity and quality of the retrieved documents completely depends on both the search queries and the search engine used.
  - Web Focused Crawling: this approach uses a specific type of program, named focused crawler. A focused crawler is a program created to retrieve data from the Web, but instead of submit multiple queries to a specific search engine, a focused crawler selectively searches for Web documents (pages) belonging to a specific topic by employing the hyperlink structure of the web, i.e. the URL.

    In detail, the Web crawling process starts with a set of pre-defined URLs. Usually, the crawler connects to a specific server or to a pre-defined set of URLs, and starts the downloading process from it. Before starting the actual crawl process, domain-specific vocabularies are semi-automatically gathered from these webpages (for all the wanted

languages). The vocabularies are very important in the process as they are used to find the seed URLs of the crawl, and consequently the "driver queries"[10] to steer the crawling process to pages that contain the wanted topic/domain. To settle a set of seed URLs for each language, the gathered vocabularies is queried in some search engine, e.g. Bing, Yahoo and Google, and the resulted URLs are used as seed URLs. Then, a priority queue that holds the URLs of the to-be-visited pages is initialised with these seed URLs. It is in this point that the actual crawl process starts. One by one, the head URL of the URL queue is removed and the page pointed to by the URL is visited. The data inside the page is extracted and the language of the page is automatically detected. If the language is one of the wanted ones the page content is matched against the driver query. If the math between the page and the driver query similarity exceeds a threshold the page content is saved and, the out-links of each fetched page are extracted, scored and prioritised according to some pre-defined rule. Then, the crawling process continues until it comes to a dead end or until some restriction defined in the crawling policy is met. The set of policies could be the maximum number of pages to crawl, the page domain, the page language, amongst others.

The benefits of this approach are that focused crawling is able to find a large proportion of relevant documents on a particular topic/domain and it is able to effectively discard irrelevant documents. By a way of example, a topical crawling approach can be used when corpora are needed to compensate for the limitations of general resources, such as general-purpose dictionaries, which do not cover vocabulary for special domains. As this approach is limited to a pre-defined topic/domain and vocabulary, it retrieves more accurate results, but compared to the previous approach requires an additional effort.

Text Formatting The resulted data could be codified in a wide variety of file formats, such as HTML (*.html*), PDF (*.pdf*), Microsoft Word (*.doc*, *.docx*), etc. For these documents to be used by a corpus management tool, they need to be converted to an acceptable format (the most widely used is plain text (*.txt*) with the character-encoding scheme ASCII or UTF-8). It is also important to take into account that some of these documents could contain information about one or more aspects of the data, such as descriptive or structural. Thus, they should be excluded from the retrieved documents, e.g. HTML *tags*. As Sinclair, 1991:21 pointed out: "*The safest policy is to keep the text as it is, unprocessed and clean of any other codes*".

Storage The last compilation stage is the data storage. Despite of the triviality of this task it is very important due to the fact that the collected data needs to be correctly identified and stored, to be, in the meantime, easily accessed. The most common way of doing this is through the use of a root directory, where the files, correctly identified are well-organised into folders and subfolders.

Representativeness This additional stage should be considered in order to determine whether the samples are representative, or not, to the object of study (cf. Lavid López, 2005).

As mentioned by Biber, 1988:246, "*the representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalisability of the results of the research*". Furthermore, he also emphasises: "*a corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language*" (Biber, 1988:246). Although he remains conscious of the difficulties involved in compiling a corpus that could be defined as representative of a particular linguistic feature (Biber, 1988), the truth is that even today the concept of representativeness is still surprisingly imprecise, considering its acceptance as a central characteristic that distinguishes a corpus from any other kind of collection. Moreover, despite some authors agree with the importance of the quality and representativeness of the samples used

[10] A driver query is a specific type of query containing the topic/domain vocabulary of a particular language.

to compile a corpus (cf. Biber, 1988; 1990; 1993; 1995; Atkins et al., 1992; Quirk, 1992; EAGLES, 1994; 1996a;b), still exists a surprising scarcity of studies devoted to analyse the quality and representativeness of a corpus, as pointed out by Flowerdale, 2004:18: "*Several corpus linguists have raised issues concerning the size and representativeness of specialised corpora as well as the generalizability of their findings. In fact, these are thorny issues which have also been widely debated in the literature on corpus studies in general, and to which there seem to be no easy answers.*".

Nevertheless, the work done by Corpas Pastor and Seghiri, 2007a;b presents for the first time a method to quantify, a posteriori, the minimum number of documents and words that should be included in a specialised language corpus. Afterwards, it is not possible to establish the minimum number of documents for a given corpus a priori because the size will always depend on the language and text types involved (Corpas Pastor and Seghiri, 2007a:171). Thus, they used the N-Cor algorithm to create an application named ReCor to measured a posteriori the corpus representativeness. This application is able to automatically determine the representativeness threshold for a given corpus. The ReCor application can be seen as a good starting point to future works on this important task – still without a consensus in the research community.

2.3.2 Comparability Degree in Comparable Corpora

One of the general characteristics of comparable corpus is the degree of comparability that documents in these corpora should have. In theory, this may seem to be fairly obvious, a comparable corpus should be balanced in quantity and quality. In practice, comparability and/or parallelism is considered an extremely complex issue.

Features Selection In order to measure the degree of comparability of a comparable corpus, a number of features need to be selected. The choice of these similarity features is influenced by different factors, such as the aim for which the comparable corpus is built for, or even the methodology employed for its acquisition. As stated by Leturia et al., 2009:55: "*the criteria to define comparability are not universal and depend on the type of comparable corpus we want and the task we want to use the corpus for*". Amongst the literature there are two types of works on comparable corpora, which induce different choices (Goeuriot et al., 2009:56): general language works (where texts of corpora usually share a domain and a period) and specialised language works (where choice of criteria is various). By way of example, a comparable corpus made of news articles will tend to be built relying publication dates, in addition to the domain and topic. Bearing this in mind, it is easy to understand that some parameters certainly have precedence over others, always depending on the purpose that the corpus is built for. In the following paragraphs some of these parameters are presented, along with a short description about their relevance to the task they are used for.

Regarding on content, Morin et al., 2007 suggest that, for the task of terminology extraction the quality of a comparable corpus might be more important than its size. Indeed, they obtained better alignment results with small corpus than previous works involving the English language (Morin et al., 2007:671). Nevertheless, the experiment was performed by using two corpora retrieved from a specialised domain, which explains why the similarity features should be based not only on the domain but also on the type of discourse. Thus, the type of discourse could be seen as a criterion to get higher levels of comparability. Goeuriot et al., 2009, apart from topic and domain, also considered the type of discourse, which proved to increase the degree of comparability between the documents. Also relying in the context of terminology extraction, Leturia et al., 2009:55 consider the domain and topic similarity more important than genre and size. Another example is given by Gamallo and López, 2010, in which the authors also relied on topic restrictions and language to gather comparable articles from Wikipedia. A complete different approach is described in Saralegi et al., 2008, in which the authors proposed to measure the comparability of a corpus by computing the semantic similarities at the document level. The hypothesis behind this is that the containment of many document pairs with a fairly high

semantic similarity improves terminology extraction based on context similarity. The assumption is that this method, somehow, is able to measure the level of comparability between pairs of documents.

As regards to extra-linguistic features, Braschler and Scäuble, 1998 took advantage of external indicators to find similarities between pairs of documents. As the same story is usually published on similar dates by news agencies, they used the publication date as an indicator to align pairs of articles (Braschler and Scäuble, 1998:185). Yet, there are other features that should be considered. When documents are extracted from the Web, the structure and the context that describes the documents origin could be retrieved to classify them. An easy way to access this information is to look at the internal HTML structure marked by HMTL tags and analyse it using, for instance regular expressions (Goeuriot et al., 2009:57). By a way of example, Goeuriot et al., 2009:56 stated that, apart from the period, the document authorship could be used, since authors sharing the same style are likely to produce similar texts.

However, most of the previous works combine both linguistic and extralinguistic criteria (cf. Braschler and Scäuble, 1998; Goeuriot et al., 2009). For instance, Bekavac et al., 2004 and Skadiņa et al., 2010b choose as parameters of comparability the domain and the topic as linguistic criteria and the size and the time span as extra-linguistic criteria. Another example can be found in Talvensaari et al., 2007 and Hashemi et al., 2010, where the document topics and their publication dates were used to align comparable documents.

In short, comparability is ensured by using several characteristics which can refer to the text creation context (publication dates, authorship, etc.), or to the text itself (topic, genre, etc.). Table 1 tries to put these features in perspective, i.e. tries to give a general idea not only about the most common features used to measure the documents comparability, but also the most frequent retrieve mechanism used to access them.

Table 1: Common similarity features used to measure the documents similarity along with the most common retrieving mechanisms.

	Similarity Features	**Retrieve Mechanism**
Linguistic	genre	words-frequency; keyword extraction; POS tagging; semantic similarity measures
	domain	
	type of discourse	
	topic	
Extra-linguistic	publication dates	regular expressions
	authorship	
	time span	
	size	

Criteria of Comparability and Parallelism One of the general characteristics of a comparable corpus is the degree of comparability that documents in these corpora should have. In theory, this may seem to be fairly obvious. A comparable corpus should be balanced in quantity and a certain quality of texts. In practice, comparability and/or parallelism is considered a complex issue, which can be applied to different levels, such as document collections, individual documents, paragraphs or even to sentences (Skadiņa et al., 2010a:8-9). Additionally, there has been no agreement on the degree of similarity that documents in comparable corpora should have, or even agreement about the criteria for measuring parallelism and/or comparability in comparable corpora. As pointed out by Sharoff, 2010:1: "*the notion of comparable corpora rests on our ability to assess the difference between corpora which are claimed to be comparable, but this activity is still art rather than proper science*". Nevertheless, there have been some

attempts to determine and specify different levels of comparability/parallelism in comparable corpora (cf. Braschler and Scäuble, 1998; Bekavac et al., 2004; Fung and Cheung, 2004; Skadiņa et al., 2010a). In the next paragraphs some of these attempts are described in detail.

Braschler and Scäuble, 1998 propose a five-level relevance scale in order to assess the quality of comparable documents alignment. The levels of relevance used to align pairs of documents are (Braschler and Scäuble, 1998:190):

i) Same story: where two documents cover exactly the same story/event.

ii) Related story: two documents deal with the same event or topic from a slightly different viewpoint. Alternatively, one of the documents may cover the same event or topic, but the topic is only a part of a broader story, or the article is composed by multiple stories.

iii) Shared aspect: two documents address various topics, but at least one of them is shared.

iv) Common terminology: the events or topics are not directly related, but they share a considerable amount of terminology.

v) Unrelated: the similarity between the documents is slight or non-existent.

Then, Bekavac et al. in 2004 introduced the notion of two levels of comparability of corpora (cf. Bekavac et al., 2004). According to the authors, these levels of comparability could be called "light" and "hard" (Bekavac et al., 2004:p.1188):

i) Corpora are said to be lightly comparable when their similarity is only in terms of extra-linguistic and extra-textual features, such as size, time-span, text genres, gender and/or age of the authors, etc.

ii) Hard comparable corpora is dependent on the previous collected "lightly" comparable corpora. In detail, this second type of comparability derives from the first one by applying certain language technology tools/techniques[11] and some pre-defined parameters of their usage, with the purpose of finding out which documents in lightly comparable corpora deal with similar topics. Then, the resulted subsets of lightly comparable corpora that have been selected by those tools/techniques can be considered as "hard" comparable corpora.

Also in 2004, Fung and Cheung, 2004 quantify the types of comparable corpora into three categories:

i) Parallel: a sentence-aligned corpus containing bilingual translations of the same document.

ii) Noisy-parallel: also called a "comparable" corpus, containing non-aligned sentences that are mostly bilingual translations of the same document, focused on the same thematic topics, with some insertions and deletions of paragraphs.

iii) Very-non-comparable: a corpus that contains far more disparate, very-non-parallel bilingual documents that could either be on the same topic (in-topic) or not (off-topic).

Another important work is presented by Skadiņa et al., 2010a, in which the authors present four levels of comparability of comparable corpora:

i) Parallel: texts considered as accurate or approximate translations with minor variations in language. They give examples of legal documents, software manuals, and fiction translations.

[11] These techniques could be: simple comparison of frequency lists of lemmas and/or collocations; named entity recognition, classification and comparison; document classification; term extraction comparison, etc. (Bekavac et al., 2004:1188).

ii) Strongly comparable: texts closely related containing the same event or describing the same subject. The given examples are: texts written by the same source, with the same editorial control, in different languages; and texts concerning the same subject, written by independent news agencies (e.g. Wikipedia articles).

iii) Weakly comparable: texts of the same narrow or broader domain and genre, but describing different events, or varying in subdomains and specific genres. An example of that is the database administrator guide for MySql in two different languages.

iv) Non-comparable: pairs of texts that do not have much in common. The Web is an example of this type of texts.

Despite the concept of comparability is considered as a complex issue, several degrees of comparability of comparable corpora were proposed by Braschler and Scäuble, 1998, Bekavac et al., 2004, Fung and Cheung, 2004 and Skadiņa et al., 2010a. Even though these four different approaches are not able to be directly compared (e.g. due to its subjectivity), table 2 put them side-by-side in order to show their relationship.

Table 2: Levels of comparability in comparable corpora presented in the literature.

<table>
<tr><th></th><th>Braschler and Scäuble, 1998</th><th>Bekavac et al., 2004</th><th>Fung and Cheung, 2004</th><th>Skadiņa et al., 2010a</th></tr>
<tr><td rowspan="5">Linguistic Criteria</td><td>-</td><td>-</td><td>Parallel</td><td>Parallel</td></tr>
<tr><td>Same story</td><td rowspan="4">Hard</td><td rowspan="4">Noisy-Parallel</td><td>Strongly comparable</td></tr>
<tr><td>Related story</td><td rowspan="3">Weakly comparable</td></tr>
<tr><td>Shared aspects</td></tr>
<tr><td>Common terminology</td></tr>
<tr><td>Extra-linguistic Criteria</td><td>Unrelated</td><td>Light</td><td>Very-non-comparable</td><td>Non-comparable</td></tr>
</table>

As table 2 shows, the comparable corpora criteria defined by Skadiņa et al., 2010a as "strongly" and "weakly" matches to the noisy-parallel and "hardly" criteria, since they share the same and/or similar topic, as "hardly" and noisy-parallel criteria do (Bekavac et al., 2004; Fung and Cheung, 2004). Moreover, the first four levels defined by Braschler and Scäuble, 1998 also fall within this range. If by on one hand Braschler and Scäuble, 1998's classification presents a greater granularity, one the other hand Bekavac et al., 2004, Fung and Cheung, 2004 and Skadiņa et al., 2010a do not make any distinction between the information specificity shared between two documents. In the lower level of comparability, Braschler and Scäuble, 1998 and Skadiņa et al., 2010a do not explicitly consider any kind of extra-linguistic features in their criteria as Bekavac et al., 2004 and Fung and Cheung, 2004 do. Finally, it is worth to notice that Fung and Cheung, 2004 and Skadiņa et al., 2010a reclaim a higher level of comparability, the parallel level, which corresponds to pairs of texts with minor variations in language.

3 Related Work

This section briefly presents a set of works that can be somehow related to this research. These works were categorised according to the approach followed by the authors, i.e using *Existing Collections*, a *Web Search Engine*, a *Web Focus Crawler* or a *Hybrid* approach (see section 3.1, 3.2, 3.3 and 3.4, respectively). As we will see in the next section there are two main approaches that can be used to gather comparable texts from the Web:

- assemble monolingual corpora independently in multiple languages using the same constraints for each language;
- or establish the constraints for one language and from the retrieved documents that fulfil these constraints generate queries for other languages using, for example Cross-Language Information Retrieval (CLIR) techniques.

3.1 Existing Collections

The initial research on comparable corpora compilation has its roots on prior corpora and news agencies collections. By taking advantage of the multilingual data available, some techniques were used to extract and align comparable documents from large collections of data, handmade or automatically created. By a way of example, the newspaper Portuguese corpus CETEMPúblico (Santos and Rocha, 2001), the Reuter's English corpus (Lewis et al., 2004), the Spanish CREA[12] corpus and the ICAME corpora (Brown, LOB, etc.) are just some of the examples of corpora already collected and publicly available for consultation (for more details see Maia, 2003).

Focusing in the purpose of producing bilingual comparable corpus, Bekavac et al., 2004 exploited two monolingual newspaper subcorpora from larger reference corpora of Bulgarian and Croatian. The main idea behind this work was the identification of common features and their further linkage, in other words, align comparable documents that are found in pre-collected monolingual corpora. The news titles, keywords and the publication date were the alignment criteria used in this work.

Also using two document collections in different languages, several research works (cf. Talvensaari et al., 2007; Hashemi et al., 2010) applied CLIR-based approach to create bilingual comparable corpora. The Cross-Language Information Retrieval (CLIR) approach broadly consists in retrieving documents written in a language different from the language of the query. In detail, firstly irrelevant vocabulary, such as stopwords[13] from the source documents is filtering out. After that, part of the remaining words (ranked based on their occurrence or not), are considered keywords, which are translated to the target language. Then, these translated keywords are used as queries to be run against the target document collection, in order to retrieve relevant documents.

Another example on how two monolingual document collections were used to create a comparable corpus is presented in Talvensaari et al., 2007. On their work, articles from a Swedish news agency and a U.S. newspaper are used to create a Swedish-English comparable corpus through a CLIR-based approach. Firstly, in order to ensure a better keywords extraction, the TWOL lemmatiser (Koskenniemi, 1983) is utilised to lemmatise inflected source document words and to decompose compound words in the Swedish collection. Relevant words in the source documents are then identified and extracted using the Relative Average Term Frequency (RATF) formula (cf. Pirkola et al., 2002). After that, the extracted keywords are translated into the target language, i.e. to English, with a dictionary-based query translation program. Then, these English translations are used as queries and ran against the target collection using the Lemur [14] retrieval system. The alignment pair is made only if the retrieved documents

[12] `http://www.rae.es`

[13] Usually a stopword list contains very rare words and very frequent words, mostly functional, like prepositions, determiners or pronouns.

[14] `www.lemurproject.org`

match a given date and a similarity score criteria. The authors took advantage of the resulting comparable corpora as a similarity thesaurus to translate queries along with a dictionary-based translator. In the end, the authors found that the combination of both approaches outperformed translation schemes where dictionary-based translations or corpus-based translations were used alone.

Also following a CLIR approach, Hashemi et al., 2010 compiled a Persian-English comparable corpus from two collections, the BBC News in English and the Hamshahri news in Persian. The applied method is simpler than the previous one (Talvensaari et al., 2007) since no pre-processing was performed on the Persian documents collection. Furthermore, instead of using a dictionary-based query translation program, to translate the keywords, they used a simple English-Persian dictionary. To deal with the translation of out of vocabulary words (e.g. proper words), they took advantage of the Google's machine translation system. Once the query in the target language is created and the documents retrieved, they used Lemur [14] to rank the documents, in this case according to their similarities to the query. It is important to mention that the similarity criteria used in this work to align the documents were the news' publication date and topic.

3.2 Web Search Engine

Corpora compilation has been energised by the recent growth of Internet, which naturally exploited new methods and approaches to compile corpora, semi-automatically. An example of that is the work done by Ghani et al., 2001, in which a corpus was built using a search engine. The general procedure consists in taking two sets of documents as input, the relevant and the non-relevant to the target language topic/domain, respectively. Given these initial documents, a term selection method is used to score and retrieve relevant an irrelevant words to the query. Then, the query is sent to a search engine and the highest ranked document is retrieved and added to the set of relevant or non-relevant documents according to the resulted filter's classification – the remaining results (hits) are stored to avoid re-querying the search engine. Next, the algorithm iterates itself by updating the set of documents and consequently creating new queries. If a repeated query appears, the next hit from the cached results is visited. The algorithm stops when all the hits have been visited. As a result, a corpus is automatically built through an automatically Web-search queries generation.

Also regarding the compilation of comparable corpora, Leturia et al., 2009 developed a search engine-based approach for acquiring specialised Basque-English comparable corpora from the Web. In this work instead of using a set of seed keywords, the authors used a sample mini-corpus to start the process. The most representative words are automatically extracted from it and a final domain-filtering step is performed using document-similarity measures with this sample corpus. In detail, the keywords are extracted by using both the Relative Frequency Ratio (RFR) (cf. Damerau, 1993) and a frequency measure named Log Likelihood Ratio (LLR). To measure the documents similarity they utilised the LLR for scoring the terms extracted from the gathered documents, and the cosine similarity measure to calculate the similarity between the gathered documents and the documents in the sample. In order to obtain a bilingual comparable corpora, the authors proposed two different variants of this method. The first one consists of using a sample mini-corpus for each language, to trigger the corpus-collection process for each language, independently. The second method only uses one sample mini-corpus in one of the languages, but requires dictionaries for translating the extracted seed words and performing the topic filtering for the other languages.

3.3 Web Focus Crawling

The amount of information available through the Web affects the performance of general-purpose search engines, which usually ignore or simple do not show some segments of the Web. By a way of example, Google's search engine uses a special algorithm to generate search results. Despite Google only shares general facts about its algorithm, as specifics are a company secret, it is

known that the company uses automated programs named spiders or crawlers, like other search engines do. Also like other search engines, these crawlers are used to build a huge index of keywords/webpage, which is then used to generate search results in milliseconds. With this in mind, it is easy to understand that a corpus created through a search engine approach will always depend on both the set of seed words and the used search engine API[15].

A solution to this constraint is the creation of a Web focused crawler. This specific type of crawler aims to selectively seek out pages that are relevant to a pre-defined set of topics. One of the first authors to present this idea was Chakrabarti et al., 1999. In this work, it is argued that a focused crawler "*seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the web*" (Chakrabarti et al., 1999:1624). Thus, a focused Web crawler can be seen as a program that aims to retrieve data from the Web, but instead of submit multiple queries to a specific search engine, it selectively searches for Web documents (pages) belonging to a specific topic, by employing the hyperlink structure of the web, i.e. the URLs. The topics are specified not using keywords, but instead seed documents. Rather than collecting and indexing all accessible webpages, a focused crawler analyses its crawl boundary to find the links that are likely to be most relevant for the crawl, avoiding this way irrelevant sections of the Web. This leads to significant savings in hardware and network resources, and helps to keep the crawl up-to-date. Nevertheless, compared to the previous approach this approach requires a tremendous effort.

Regarding the literature in this topic, Chakrabarti et al., 1999, Stamatakis et al., 2003, Pirkola, 2007, Talvensaari et al., 2008 and Skadiņa et al., 2010b are only some of the authors that used a focused Web crawler to exploit corpora from the the Web. For example, focusing in the purpose of exploit multilingual comparable documents from the Web, Talvensaari et al., 2008 used this approach to gather genomics-specific text in English, Spanish and German. Before the crawling operation, domain specific vocabulary was collected separately in all three languages and used to acquire relevant seed URLs. The selected URLs were then employed, as driver queries, in the crawling phase to identify relevant pages, from which text paragraphs were then extracted. The language of each paragraph is detected with a simple n-gram-based algorithm (Cavnar and Trenkle, 1994). After that, if the paragraph matches with one of the considered languages, it is matched against the driver query of the particular language. These queries contained approximately 300 domain words. But, the paragraph is only saved to the hard-drive if the match score exceeded a specific threshold. This score is calculated based on the proportion of domain words (words included in the driver query) versus the total number of words in the extracted paragraph. Moreover, the URLs found are kept in memory in order to be visited only once. The same technique is applied to paragraphs because often the same paragraph appears on different webpages (or different URLs point to pages with the same content).

3.4 Hybrid

Huang et al., 2010 propose a method for assembling a comparable corpus from Chinese-English news collections. While the previous works rely mainly on a focused crawling approach to compile comparable documents, Huang et al., 2010 use crawling in addition to the CLIR-based approach in order to ensure higher comparability between document pairs. Firstly, they harvested the original source and target document sets from news websites using an open-source crawler. Then, the keywords were extracted from the source documents. The approach used to extract the most relevant keywords is mainly based on the extraction of multi-word expressions (MWE) followed by word ranking method. More precisely, the MWEs are extracted using the LocalMaxs selection algorithm and ranked by a relevance measure proposed by Silva et al., 1999, the Symmetric Conditional Probability measure. Single word candidates are identified and ranked with Term Frequency-Inverse Document Frequency (TF-IDF). In the next step, the extracted keywords are translated into the target language by a bilingual dictionary, in order to be used in the

[15] Application Programming Interface.

filtering and retrieval process. Specifically, the similarity criteria used to filter and mapping the correlations between source and target documents were: the news' publication date; a similarity score provided by the Lemur [14] retrieval system; and the Keyword Similarity between Document pairs (KSD) measure. In the end, the authors show that their approach is effective to mine Chinese-English document pairs.

4 Existing Corpora Compilation Solutions

The World Wide Web has become a primary meeting place for information and recreation, for communication and commerce. Millions of users have created billions of webpages in which they expressed their vision about the world. This linguistic and cultural content is considered a golden mine for lexicographers, linguists, translators, teachers and other language professionals. As a source of machine-readable texts for corpus linguists and researchers in complementary fields like Natural Language Processing (NLP), Information Retrieval (IR) and Text Mining for example, the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness. To this end, several tools (i.e. web crawlers, language identifiers, HTML parsers, HTML cleaners, etc.) have been developed and combined in order to produce corpora from this golden mine. Bearing this in mind, this section aims to review the most relevant approaches/ methodologies/ tools capable of compiling parallel and comparable corpora from the Web (section 4.1 and 4.2, respectively).

4.1 Mining Parallel Corpora

It is already a fact that the Internet can be seen as a large multilingual corpus due to its huge number of multilingual websites, in which different pages can contain the same written text in different languages. This means that some of their webpages can be paired into *bitexts* (or parallel texts).

Bitexts have become a very important source of knowledge, specially for the Machine Translation (MT). Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) are just some examples of MT sub-areas where this kind of resource is fundamental, e.g. for the process of training (Hutchins and Somers, 1992). In fact, there are corpora which have been obtained from the Internet with the purpose of training SMT. The most known example is the Europarl Corpus (Koehn, 2005). Another example of corpus-based MT can be found in Caseli and Nunes, 2007 and Sánchez Martínez and Forcada, 2009, where the authors created Rule-Based Machine Translation (RBMT) systems by extracting translation rules from parallel corpora.

Nevertheless, the problem of collecting these data presupposes a significant technical challenge and the question remains: How to find these parallel texts and obtain an aligned parallel corpus from them? Bearing this in mind, different systems have been developed to harvest *bitexts* from the Internet. The next sections try to present and describe the most relevant and important systems developed so far for this task.

4.1.1 STRAND

STRAND[16] (Structural Translation Recognition, Acquiring Natural Data) (Resnik, 1998; 1999; Resnik and Smith, 2003) can be considered as one of the earliest core Web-mining architectures capable of identify webpages which are candidates to be *bitexts*. In order to do this, it uses the structural features of documents, a content-based measure of translational equivalence, and the Web as a source for mining *bitexts* on a large scale. The general procedure includes three main steps: 1) locate possibly parallel webpages; 2) generate candidates pairs of parallel webpages; and, finally, 3) apply structural filters to the candidate set.

Locate possibly parallel webpages STRAND architecture uses AltaVista[17] search engine's advanced search to identify webpages that match specific patterns. The system looks for two types of webpages: *parents* and *siblings*. A parent page is a page that contains hyperlinks to different language versions of a document. To find these parent pages, STRAND uses language-specific boolean search queries, like `(anchor:‘‘english’’ OR anchor:‘‘anglais’’) AND (anchor:‘‘french’’ OR anchor:‘‘francais’’)`. Then, it applies simple regular expressions

[16] `http://www.umiacs.umd.edu/~resnik/strand/`

[17] AltaVista, first launched in December 1995, was one the world's first search engines and was shut down on 8 July, 2013.

to filter candidates pages, which is done by looking for webpages with language links close to each other. Sibling documents are webpages that include links to translations of the same page. Once again, boolean search queries, such as `(anchor:‘‘english’’ OR anchor:‘‘anglais’’)` are used to obtain a candidate list of such webpages. A later version of STRAND includes a web spider (i.e. a Web focused crawler) component for locating pages that might have translations. Similar to other systems (see section 4.1.3), like BITS (Bilingual Internet Text Search) (Ma and Liberman, 1999) and PTMiner (Parallel Text Mining Algorithm) (Chen and Nie, 2000), STRAND also uses Levenshtein distance (Levenshtein, 1966) to compute the difference between URLs. In detail, URLs pairs are consider candidate pairs if only differ in a prefix or suffix. In order to identify this pre-defined prefix and suffix, they use patterns like `(.el | .en | _en | ...)`, which are manually configured.

Generate candidates pairs of parallel webpages This step uses the *text-language comparison*, *URL-matching* and the *document lengths* as features to generate candidate pairs. When the parent or sibling pages are already generated, the system reduces the set of candidate page pairs by throwing away pages that do not match the two languages of interest (text-language comparison step). Then, an additional URL-matching stage is performed. Assuming that the directory structure, on mostly of the webpages, reflects parallel organisation, the system uses a list of substitution rules (manually created) to generate new URLs. These new URLs are then matched with the list of pages, already retrieved in the other language of interest. If such URL is found, the pair with similar URLs is added to the list of candidate document pairs. This idea can be illustrated with the example given in Resnik and Smith, 2003:4: supposing that an English-Chinese website contains a page with the following URL `http://mysite.com/english/home_en.html`, on which one combination of substitutions might produce the URL `http://mysite.com/big5/home_ch.html`, the original webpage and the produced URL are probably worth considering as a likely candidate pair. The third criterion for matching uses the document lengths. Assuming that texts that are translations of one another tend to be similar in length, STRAND uses a document length filter (Smith, 2001) to reduce the size of the search space for candidates pairs.

Apply structural filters to the candidate set The final step in STRAND consists in applying structural filters to all documents pairs collected in the previous stages. To do this, STRAND assumes that translated documents use the same or a very similar HTML tag structure (markups). The system uses a markup analyser that produces a linear sequence of tokens representing tags and chunk lengths. This analyser produces tokens for: HTML start-tags `[START:label]`, representing any <label>; tokens for the HTML end-tags `[END:label]` representing </label>; and tokens for characters data `[Chunk:length]`, which represents any text of length between two HTML tags. These token sequences are then compared for every document pair. First, the system aligns them using standard sequence comparison algorithms (Hunt and McIlroy, 1976). Then, STRAND computes four values that characterise the quality of the alignment: 1) the percentage of differences between the two sequences; 2) the number of aligned nonmarkup text chunks of unequal length; 3) the length correlation of aligned non-markup chunks; 4) the significance level of that correlation. Thresholds on all values can be manually configured to perform alignment decisions. In this way, precision and recall of the extraction process can be balanced according to one's needs. Apart from manual settings, Resnik and Smith, 2003 also discuss the use of a supervised machine learning techniques. In the latter, they applied a decision tree to improve the classification performance of candidate pairs.

4.1.2 Bitextor

Bitextor[18,19] (Esplà Gomis, 2009; Esplà Gomis and Forcada, 2009; 2010) is a free/open-source application created for Unix platforms, which aims to generate translation memories using

[18] `http://bitextor.sourceforge.net`
[19] `http://sourceforge.net/projects/bitextor`

multilingual websites as a corpus source. This tool was created to be as adaptable as possible when retrieving multilingual data from any kind of websites and work with any pairs of languages. To do that, it combines context-based and URL-based heuristics to harvest aligned *bitexts* from multilingual websites.

The Bitextor workflow can be divided into three main steps: 1) downloading, processing and choosing the parameters for the comparison; 2) webpage comparison; and, finally, 3) aligning the obtained webpages. It is important to mention that Bitextor is based on two main assumptions: parallel pages should be under the same domain and they should have similar HTML structure.

Downloading, processing and choosing the parameters for the comparison In order to download the entire website, i.e. all the HTML files from a multilingual website, in a directory tree structure, Bitextor uses the HTTrack[20] application. Then, the preprocessing stage uses the Tidy[21] library to standardise invalid HTML files into a valid XHTML format, ensuring this way that the tags structure is correct. The original character encoding is also converted to UTF-8. Once the files have been download and preprocessed, some additional information is extracted from the files, such as *surface features*, *webpage content* and *URL*. Surface features, like *text-language comparison* (for this purpose Bitextor uses LibTextCat[22]), *file size ratio* and *total text length difference* are used as a indicator to discard very unlike pairs of files. Then, two element in the content of the webpages are considered as parameters for the comparison: the *HTML tag structure* and the *block length*. Similar to the STRAND approach, Bitextor assumes that two parallel webpages have the same HTML tag structure or at least a similar one. Therefore, Bitextor starts by removing all the irrelevant information within the pages, such as comments, the heading of the web page, tag parameters, irrelevant HTML tags and the extra spaces in the text blocks. In the second step, Bitextor encodes the remaining information into a string containing the *tag names* and the *text block lengths* (measured in characters). This encoded string will act as a *fingerprint* of the webpage content, which will be used as a comparison parameter when measuring the similarity between two pages. Different from other approaches, Bitextor does not download the candidate files by using rules of detection and substitution of language markers in the URL (Nie et al., 1999). Rather, Bitextor first downloads the whole website and then uses the URLs, either as one more parameter to discard or associate pairs of files. To do that, Bitextor divides the URL into three sections: *directory path*, *file name* and *variables*, which are then used to compare each section separately.

Webpage comparison process In order to compare the retrieved webpages, Bitextor compares one by one all the extracted features. Firstly, the *surface features* are compared in order to discard the most obviously incorrect pairs of files. Then, the following two methods are applied to the remaining files: a) *URLs comparison*: candidate pairs can have one difference in their URL (the file name, the variable or the directory path); b) *Webpage content comparison*: finally, those pairs that have not have been discarded so far are compared through their fingerprint. This is done using the Levenshtein edit distance algorithm (Levenshtein, 1966). For more details about the comparison methods used by the Bitextor system see Esplà Gomis and Forcada, 2010.

Aligning the obtained webpages The last step is the candidates alignment and generation of the translation memories in a TMX (Translation Memory eXchange) format. The process is based on level comparisons. Although parallel texts are usually in the same level or in very close levels, the users can limit the depth difference in the directory tree when performing the comparison. This tuning ensures that it is not necessary to compare all the files between each other, but instead only compare files within the defined level interval. In order to align a pair of XHTML files and generate the translation memory in a TMX format, Bitextor uses

[20] `http://www.httrack.com`
[21] `http://tidy.sourceforge.net`
[22] `http://software.wise-guys.nl/libtextcat`

the LibTagAligner[23] library. The method used by this library is similar to the one used by Bitextor during the comparison phase. LibTagAligner encodes the file with a fingerprint (as Bitextor does), but it uses a more detailed weight structure with the edit-distance algorithm. Once again, the user can define weights for the operations performed by the edit-distance algorithm. A detailed description about the comparison process and the alignment can be found in Esplà Gomis and Forcada, 2009 and Esplà Gomis and Forcada, 2010.

4.1.3 Other Systems

Several other systems for compiling parallel corpora can also be described as operating within the same three-stage framework as those mentioned in the previous sections. This section describes three of them (BITS, PTMiner and WeBiTex).

BITS (Bilingual Internet Text Search) (Ma and Liberman, 1999) is a Web mining system that aims to harvest multilingual texts over the World Wide Web. The BITS architecture is a simple pipeline. BITS starts with a specific pair of languages of interest and a given list of domains to search for parallel text. Then, for each website on the list, the website language identifier detects the language of the website. If it is not a multilingual website, then the system processes the next website on the list. If it is a multilingual website, the system downloads all the pages from the website recursively using the GNU Wget utility. Afterwards, the pages are converted to plain text files. Next, a language identifier detects the language of each text file. And, finally, a translation pairs finder finds all the translation pairs and stores them in a database. To find the translation pairs, BITS uses a large bilingual dictionary and n-grams to compute a content-based similarity score. If the resulted score is lower than a pre-defined threshold these pages are not considered. Many details about the techniques used are left unspecified in Ma and Liberman, 1999, such as the dictionary entries matching, the threshold for the similarity score and the distance threshold. In addition to cross-lingual lexical matching, the system filters out candidate pairs that do not match well in terms of file size, document content (numbers, acronyms and some named entities) or paragraph counts.

PTMiner (Parallel Text Miner) (Chen and Nie, 2000) is a parallel text mining system that aims to automatically find parallel texts on the Web. The system starts by querying existing Web search engines for pages in a given language that may contain links to webpages that are likely to be in the other language of interest. Once these websites are located, PTMiner crawls exhaustively all the pages. In order to generate candidate pairs, the system uses a URL-matching process, similar to the one used by STRAND. Although it uses a language-specific prefixes and suffixes during the matching process as STRAND, PTMiner does not handle cases in which URL matching requires multiple substitutions. Similar to other systems, PTMiner also uses the file size, language and character set of each page as a parameter to filter out non-parallel candidate pairs. During the last stage, PTMiner applies another filter to clean the extracted corpus (Nie and Cai, 2001).

WeBiTex[24] (Désilets et al., 2008) is a free and open-source web-based systems built with the purpose of helping translators resolve typical translation problems that they encounter in their work. WeBiText allows to perform searches in a corpus comprised of over 10 million webpages from the Government of Canada and other bilingual or multilingual websites. It allows the user to search a large collection of pre-indexed sites or to launch a custom query in any of the 30 supported languages by adding a multilingual site of their choice. Search results are displayed in a split window, source and target segments. Each result features links to source and target files (to view full context), as well as the *bitext* (segmented at the sentence level). New bilingual and multilingual web sites are regularly added and the index is automatically updated with each query, thus enabling users always to get up-to-date results.

[23] `http://sourceforge.net/projects/tag-aligner`
[24] `http://www.webitext.com`

4.2 Mining Comparable Corpora

There is a growing literature on using the Web for constructing various types of text collections, including domain-specific monolingual, bilingual and multilingual comparable corpora (cf. Costa et al., 2014; 2015). Particularly, in translation their benefits have been demonstrated by several authors (cf. Bowker and Pearson, 2002; Bowker, 2002; Zanettin et al., 2003; Corpas Pastor and Seghiri, 2009). One potential solution to the lack of sufficient/up-to-date parallel corpora and linguistic resources for narrow domains is the exploitation of non-parallel mono-/bi- and multilingual text resources, also known as comparable corpora (i.e. corpora that include similar types of original texts in one or more language using the same design criteria (cf. EAGLES, 1996; Corpas Pastor, 2001:158)). Although the process of compiling comparable corpora can be manually performed (see section 2.3 for more details), nowadays, specialised tools can be used to automate this tedious task. This section presents and describes in detail how the two most known tools on the market exploit corpora mined from the Web.

4.2.1 BooTCaT

BootCaT[25] (Baroni and Bernardini, 2004) is a semi-automatic compilation program that makes use of online information to construct a Web-based corpus. The process is very simple and only requires a set of seed terms as input. Then, these seeds are randomly grouped to form tuples (i.e. a variety of combinations of the seeds), which are submitted as search query strings to the Bing[26] search engine. Additionally, BootCaT allows the user to insert, before starting the retrieval process, a list of "black" and "white" words. If on one hand, the blacklist is used to remove documents containing more than a certain number of words. On the other hand, the whitelist is used to include documents that contain a ratio, between the words in the whitelist and the total words in the document, above a certain threshold. Then, during the download process, the top n pages returned for each query are retrieved and formatted as plain text. BootCaT provides access to large amounts of data in a few minutes, reducing the time of manual intervention in the compilation process. It is also possible to build a larger corpus by repeating the process using more seeds, or even create a comparable corpus by repeating the process using similar keywords in different languages. Having this in mind, this tool is not based on automatic but semi-automatic search.

Despite of the multiple advantages, BootCaT has a few limitations, which restricts the "natural process" that is usually used to compile bi- or multilingual comparable corpora. The following paragraphs summarise some of the limitations pointed out in Baroni and Bernardini, 2004:1313 and Gutiérrez Florido et al., 2013:3.

- lack of technical support, apart from the FAQs section there is no technical documentation available;
- the searches performed in the Web only uses the boolean operator "AND", which consequently leads to less accurate searches than if boolean operators such as "NOT" and "NOR" were used;
- in order to obtain results the seed words need to be semantically related to each other, if not, the retrieved documents will not have an acceptable quality;
- despite the possibility of choosing the lengths of the tuples, the tool restricts the possible combinations of the tuple's length, i.e. it is not possible to combine tuples with length of two and three at same time, for example;
- as reported by Gutiérrez Florido et al., 2013:3, sometimes the tool freezes during the search process, which may be due to: poor selection of keywords; a poor choice of URLs; the limit

[25] `http://bootcat.sslmit.unibo.it`
[26] `http://www.bing.com`

of searches per month[27]; or even due to internal problems;

- finally, the tool does not allow to perform a new compilation without closing and opening the tool again.

Despite some drawbacks, this tool can be seen as a viable source of "disposal" corpora (Varantola, 2003) built virtually for several purposes, such as translation tasks, construction of terminologies databases and domain-specific Machine Learning tasks (Baroni and Bernardini, 2004).

BootCaT toolkit can either be used in the form of a library (a suite of Perl scripts), or used as a graphical interface. The BootCaT graphical interface is a wizard that guides the user through the process of creating a Web corpus. It is important to mention that the interface does not support all the features available in the command-line scripts. BootCaT is free and open source. In detail, the BootCaT front-end is a free software, developed in Java, that can be redistribute and/or modified under the terms of the GNU General Public License[28]. Regarding the BootCaT command-line scripts suite, it can be copied or redistributed under the same terms as Perl[29].

4.2.2 WebBooTCat

Sketch Engine[30] (Kilgarriff et al., 2004) is a leading corpus query tool. Apart from offering a corpus-building tool, it also provides access to corpora online and several analysis tools in a single platform. Nevertheless, the most relevant tool for this work is the WebBootCaT[31] (Baroni et al., 2006), which allows the user to create a specialised corpus from the Web in a few minutes. To do that, it only requires a set of seed words or URLs as input, describing the domain of investigation. This tool can be seen as a Web-service version of the BootCaT tool, but rather than download and install a software, WebBootCaT has the advantage of been already installed on a Web server. Yet, this tool is only freely available on a trial basis or through the commercial product subscription.

[27] BootCaT uses the Bing search engine to find webpages relevant to the domain. In order to perform this automated task BootCaT requires an account key from Bing, which has limits in the number of queries that can be submitted per month.

[28] `http://www.gnu.org/licenses/gpl.html`

[29] `http://dev.perl.org/licenses/artistic.html`

[30] `http://sketchengine.co.uk`

[31] `https://www.sketchengine.co.uk/documentation/wiki/Website/Features#WebBootCat`

5 iCorpora: Compiling, Managing and Exploring Multilingual Data

As shown in the previous section, several semi-automatic compilation tools have been proposed so far, either capable of exploiting comparable or parallel corpora from the Web. Nevertheless, these compilation tools are scarce or proprietary, simplistic with limited features or too complex to be used by ordinary people's. Moreover, regarding comparable compilation tools, they were built to compile one monolingual corpus at a time and do not cover the entire compilation process (i.e. apart from compiling monolingual comparable corpora, they do not allow managing and exploring both parallel and multilingual comparable corpora). Thus, their simplicity, lack of features, performance issues and usability problems result in a pressing need to design new compilation tools tailored to fulfil not only translators' and interpreters' needs, but also professionals' and ordinary people's.

Departing from a careful analysis of the weaknesses and strengths of the current compilation solutions, we started by designing and developing a robust and agile web-based application prototype to semi-automatically compile, manage and explore both parallel and multilingual comparable corpora, which we named *iCorpora*. In detail, iCorpora will aggregate three applications: *iCompileCorpora*, *iManageCorpora* and *iExploreCorpora* (section 5.1, 5.2 and 5.3, respectively).

5.1 iCompileCorpora

The first application, iCompileCorpora can be simply described as a web graphical interface that will guide the user through the entire corpus compilation process. Designed and implemented from scratch, this application aims to cater to both novice and experts in the field. It will not only provide a simple interface with simplified steps, but also will permit experienced users to set advanced compilation options during the process. iCompileCorpora will allow the user to compile both comparable and parallel corpora.

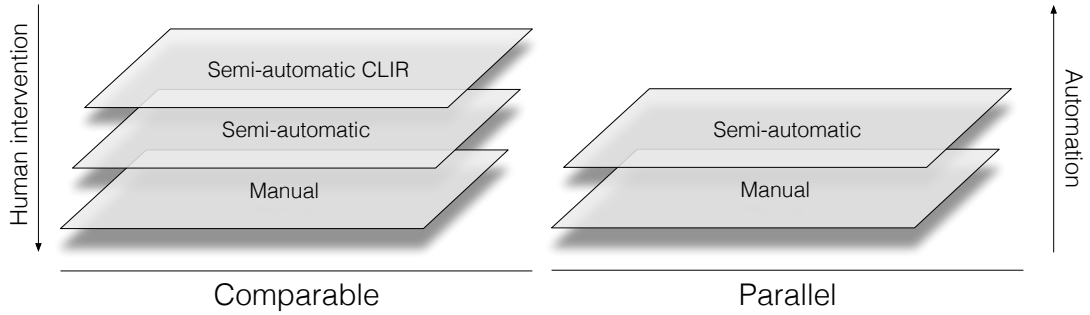

Figure 1: iCompileCorpora layered model.

Compiling Comparable Corpora The dimensions that comprise iCompileCorpora can be represented in a layered model comprising a manual, a semi-automatic web-based and a semi-automatic Cross-Language Information Retrieval (CLIR) layer (see Figure 1). This design option will permit not only to increase the flexibility and robustness of the compilation process, but will also hierarchically extend the manual layer features to the semi-automatic web-based layer and then to the semi-automatic CLIR layer. In detail, the manual layer represents the option of compiling monolingual and multilingual comparable corpora. It will allow for the manual upload of documents from a local or remote directory onto the platform. The second layer will permit the exploitation of both mono- and multilingual comparable corpora mined from the Internet. Although this layer can be considered similar to the approaches used by BootCaT and WebBootCat (see section 4.2), it has been designed to address some of their limitations (e.g. allow the use of more than one boolean operator when creating search query strings), and to improve the User Experience (UX) with this type of software. As nowadays there is an increasing

demand for systems that can somehow cross the language boundaries by retrieving information in various languages with just one query, the third layer aims to answer this demand by taking advantage of CLIR techniques to find relevant information written in a language different to the one semi-automatically retrieved by the methodology used in the previous layer.

Compiling Parallel Corpora Regarding the parallel compilation process, iCompileCorpora will also allow for the manual upload of parallel documents from a local or remote directory onto the platform (see Figure 1, manual layer). The second layer, i.e. the semi-automatic layer will permit the exploitation of parallel corpora mined from the Web. As shown in section 4.1, acquiring parallel data involves several tasks, such as crawl the web, parse the structure of each fetched webpage and extract its metadata, link ranking, cleaning, text classification, near-duplicates removal, etc. Bearing this in mind, efficient focused web crawlers can be built by adapting existing open-source frameworks like Heritrix[32], Nutch[33] and Bixo[34]. Search engine Application Programming Interfaces (APIs) can also be used to identify in-domain webpages (Hong et al., 2010) or multilingual web sites (Resnik and Smith, 2003). By a way of example, Almeida and Simões, 2010 describe a simple approach to detect which links point to translations. At this point is not yet clear which approach/ algorithms and/or frameworks iCompileCorpora will used. Nevertheless, the methodology proposed by Resnik, 1998; 1999; Resnik and Smith, 2003 seems to be the most appropriated, i.e. locate possibly parallel webpages, generate candidates pairs of parallel webpages, and then apply structural filters to the candidate set in order to clean "noisy data". By taking advantage of existing open-source frameworks and APIs, iCompileCorpora aims to be a modular and user-friendly web application with simple configuration steps.

5.2 iManageCorpora

The second application is called iManageCorpora (see Figure 2) This application will be specially designed to: manage (i.e. it will allow to edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as to manage corpora into domains and sub-domains); measure the similarity between documents; and to explore the representativeness of the corpora (cf. Corpas Pastor and Seghiri, 2009).

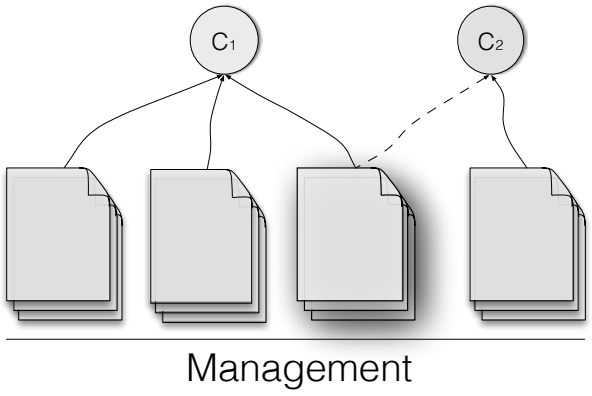

Figure 2: iManageCorpora layered model.

5.3 iExploreCorpora

Finally, iExploreCorpora intends to offer a set of concordance features, such as search for words in context, automatic extraction of the most frequent words and multi-words, amongst other features.

[32] http://crawler.archive.org/
[33] `http://nutch.apache.org`
[34] `http://openbixo.org/`

Figure 3: iExploreCorpora layered model.

6 Concluding Remarks

This report starts by presenting a comprehensive review on fundamental concepts related with corpus linguistics, concepts of corpus linguistics and corpus compilation. Various works that can be somehow related to this research were described in detail and a careful analysis of the weaknesses and strengths of the current compilation solutions on the market was also performed. Then, based on the findings, a new and agile web-based application prototype to semi-automatically compile, manage and explore both parallel and multilingual comparable corpora was proposed and described in detail. This application, named iCorpora aims to help not only translators and interpreters, but also researchers working with TMs, EBMT and SMT systems. The proposed system will be composed by three applications: iCompileCorpora, iManageCorpora and iExploreCorpora. The first application aims to guide the user through the entire corpus compilation process, either parallel or comparable. It will not only provide a simple interface with simplified steps, but also will permit experienced users to set advanced compilation options during the compilation process. The second application, called iManageCorpora will be used to: manage (i.e. it will allow to edit, copy and paste sentences and documents from and to documents and corpora respectively, as well as to manage corpora into domains and sub-domains); measure the similarity between documents; and to explore the representativeness of the corpora. The third application, named iExploreCorpora intends to offer a set of concordance features, such as search for words in context, automatic extraction of the most frequent words and multi-words, amongst other features.

References

Aarts, J. (1991). Intuition-Based and Observation-Based Grammars. In Aijmer, K. and Altenberg, B., editors, *English corpus linguistics: Studies in honor of Jan Svartvik*, pages 44–62.

Aarts, J. and Meijs, W. (1984). *Corpus Linguistics: Recent developments in the use of computer corpora in English language research.* Number V. 1 in Costerus. Rodopi.

Abaitua, J. (2002). Tratamiento de corpora bilingües. *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita. Barcelona: Fundación Duques de Soria-Edicions Universitat de Barcelona (Manuals UB, 53)*, pages 61–90.

Abercrombie, D. (1965). *Studies Phonetics and Linguistics.* Oxford University Press.

Almeida, J. J. o. and Simões, A. (2010). Automatic Parallel Corpora and Bilingual Terminology extraction from Parallel WebSites. In *3rd Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 50–55, Valletta, Malta.

Atkins, S., Clear, J., and Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, 7(1):1–16.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In M. Baker, M. F. and Tognini-Bonelli, E., editors, *Text and Technology. In Honour of John Sinclair*, pages 233–250, Philadelphia/Amsterdam. John Benjamins Publishing Company.

Baker, M. (1995). Corpora in Translation Studies. An Overview and Suggestions for Future Research. *Target*, 7(2):223–243.

Baker, M. (1996). *Corpus-based translation studies: The challenges that lie ahead*, pages 175–186. John Benjamins Publishing Company.

Baker, P. (2010). *Sociolinguistic and Corpus Linguistics.* Edinburgh University Press, Edinburgh, UK.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *4th Int. Conf. on Language Resources and Evaluation*, LREC'04, pages 1313–1316.

Baroni, M., Kilgarriff, A., Pomikálek, J., and Rychlý, P. (2006). WebBootCaT: instant domain-specific corpora to support human translators. In *11th Annual Conf. of the European Association for Machine Translation*, EAMT'06, pages 247–252, Oslo, Norway. The Norwegian National LOGON Consortium and The Deparments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway).

Behrens, H. (2008). *Corpora in Language Acquisition Research: History, Methods, Perspectives.* Trends in Language Acquisition Research. John Benjamins Publishing Company.

Bekavac, B., Osenova, P., Simov, K., and Tadić, M. (2004). Making Monolingual Corpora Comparable: a Case Study of Bulgarian and Croatian. In *4th Language Resources and Evaluation Conference*, LREC'04, pages 1187–1190, Lisbon, Portugal.

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press, Cambridge, UK.

Biber, D. (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variations. *Literary and Linguistic Computing*, 5(4):257–269.

Biber, D. (1993). Representativiness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison.* Cambridge University Press, Cambridge, UK.

Bloomfield, L. (1933). *Language.* University of Chicago Press.

Boas, F. (1911). *Handbook of American Indian Languages*. Number Vol. 1 in Bureau of American Ethnology, Bulletin 40. U.S. Government Printing Office, Washington.

Botley, S., Glass, J., McEnery, T., and Wilson, A., editors (1996). *Proceedings of Teaching and Language Corpora*. UCREL.

Bowker, L. (2002). *Computer-aided Translation Technology: A Practical Introduction*. Didactics of translation series. University of Ottawa Press.

Bowker, L. and Pearson, J. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. Routledge.

Braschler, M. and Scäuble, P. (1998). Multilingual Information Retrieval Based on Document Alignment Techniques. In *2^{nd} European Conf. on Research and Advanced Technology for Digital Libraries*, pages 183–197. Springer.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Caseli, H. and Nunes, M. (2007). Automatic Induction of Bilingual Lexicons for Machine Translation. *Journal of Translation*, 19(1):29–43.

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused Crawling: A New Approach to Topic-specific Web Resource Discovery. *Computer Networks*, 31(11-16):1623–1640.

Chen, J. and Nie, J.-Y. (2000). Automatic construction of parallel English-Chinese corpus for cross-language information retrieval. In *6^{th} Conf. on Applied Natural Language Processing*, pages 21–28.

Chesterman, A. (2004). *Beyond the Particular*, pages 33–49. John Benjamins Publishing Company.

Chomsky, N. (1955). The Logical Structure of Linguistic Theory. A typescript chomsky wrote in preparation for his phd thesis, including hand-written notes made in preparation for the 1975 book., Massachusetts Institute of Technology. `http://alpha-leonis.lids.mit.edu/chomsky/`.

Chomsky, N. (1957). *Syntactic Structures*. Janua linguarum: Series minor. Mouton.

Chomsky, N. (1962). Paper given at Third Texas Conference on Problems of Linguistic Analysis in English, 1958, p.159.

Chomsky, N. (1965). *Cartesian Linguistics*. Harper and Row.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. Springer.

Chomsky, N. (1986). *Cartesian Linguistics: A Chapter in the History of Rationalist Thought*. University Press of America.

Chomsky, N. (1993). *Language and Thought*. MOYER BELL Limited.

Corpas Pastor, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS, Revista de Traductología*, 5(1):155–184.

Corpas Pastor, G. (2002). Traducir con corpus: de la teoría a la práctica. In *Texto, Terminología y Traducción*, pages 189–226, Salamanca, Spain. Joaquín García Palacios y María Teresa Fuentes Morán.

Corpas Pastor, G. (2003). Turicor: Compilación de un corpus de contratos turísticos (alemán, español, inglés, italiano) para la generación textual multilingüe y la traducción jurídica. In *Panorama actual de la investigación en traducción e interpretación*, pages 373–384, Granada, Spain. Atrio.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Studien Zur Romanischen Sprachwissenschaft Und Interkulturellen Kommunikation. Peter Lang Pub Incorporated.

Corpas Pastor, G. and Seghiri, M. (2007a). Determinación del Umbral de Representatividad de un Corpus mediante el Algoritmo N-Cor. *SEPLN: Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 39:165–172.

Corpas Pastor, G. and Seghiri, M. (2007b). Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness. *Translation Journal*, 11(3):19.

Corpas Pastor, G. and Seghiri, M. (2009). Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). In Beeby, A., Inés, P., and Sánchez-Gijón, P., editors, *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*, Benjamins translation library, chapter 5, pages 75–107. John Benjamins Publishing Company.

Costa, H., Corpas Pastor, G., and Seghiri, M. (2014). iCompileCorpora: A Web-based Application to Semi-automatically Compile Multilingual Comparable Corpora. In *Translating and the Computer 36 - AsLing*, London, UK.

Costa, H., Corpas Pastor, G., Seghiri, M., and Mitkov, R. (2015). (*Accepted*) iCorpora: Compiling, Managing and Exploring Multilingual Data. In 7^{th} *Int. Conf. of the Iberian Association of Translation and Interpreting Studies*, AIETI, pages 74–76, Malaga, Spain.

Cruse, D. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Damerau, F. J. (1993). Generating and Evaluating Domain-oriented Multi-word Terms from Texts. *Information Processing and Management*, 29(4):433–447.

Désilets, A., Farley, B., Stojanovic, M., and Patenaude, G. (2008). WeBiText: Building Large Heterogeneous Translation Memories from Parallel Web Content. In *Translating and the Computer 30*, London, UK.

EAGLES (1994). Corpus Typology: A framework for classification. Tech Report N.2.1 written by John M. Sinclair, EAGLES Document 080294, Corpus Linguistics Group, Universidad de Birmingham, UK.

EAGLES (1996a). Preliminary Recommendations on Corpus Typology. Technical report, EAGLES Document EAG-TCWG-CTYP/P. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

EAGLES (1996b). Text Corpora Working Group Reading Guide. Technical report, EAGLES Document EAG-TCWG-FR-2. `http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html`.

Esplà Gomis, M. (2009). Bitextor, un cosechador automático de memorias de traducción a partir de sitios web multilingües. *Procesamiento del Lenguaje Natural*, 43(1):365–366.

Esplà Gomis, M. and Forcada, M. (2009). Bitextor, a free/open-source software to harvest translation memories from multilingual websites. In *Workshop Beyond Translation Memories: New Tools for Translators MT*, Ottawa, Ontario, Canada.

Esplà Gomis, M. and Forcada, M. (2010). Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93(1):77–86.

Firth, J. R. (1935). The Technique of Semantics. In *Transactions of the Philological Society*, pages 36–72.

Firth, J. R. (1957a). *A Synopsis of Linguistic Theory, 1930-1955*, pages 1–32. Blackwell, Oxford.

Firth, J. R. (1957b). *Papers in Linguistics 1934-1951*. Oxford University Press, London, UK.

Flowerdale, L. (2004). The argument for using English specialised corpora to un academic and professional language. In Connor, U. and Upton, T., editors, *Discourse In The*

Professions: Perspectives From Corpus Linguistics, pages 11–33, Amsterdam/Philadelphia. John Benjamins.

Francis, W. N. and Kučera, H. (1979). Manual of Information to accompany a Standard Sample of Present-day Edited American English, for use with Digital Computers. Original ed. in 1964 and revised in 1971. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island. `http://icame.uib.no/brown/bcm.html`.

Fries, C. (1952). *The structure of English: an introduction to the construction of English sentences*. Harcourt, Brace.

Fung, P. and Cheung, P. (2004). Mining Very-Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and EM. In *Conf. on Empirical Methods in Natural Language Processing*, EMNLP'04, pages 57–63, Barcelona, Spain.

Gamallo, P. and López, I. G. (2010). Wikipedia as a multilingual source of comparable corpora. In *3rd Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 21–25, Valletta, Malta.

Ghani, R., Jones, R., and Mladenić, D. (2001). Mining the Web to Create Minority Language Corpora. In *10th Int. Conf. on Information and Knowledge Management*, CIKM'01, pages 279–286. ACM.

Goeuriot, L., Morin, E., and Daille, B. (2009). Compilation of Specialized Comparable Corpora in French and Japanese. In *2nd Workshop on Building and Using Comparable Corpora (BUCC'09)*, pages 55–63, Singapore. ACL.

Gries, S. T. (2008). *Corpus-based methods in analyses of SLA data*, pages 406–431. Routledge, NY, USA.

Gries, S. T. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, 3(5):1225–1241.

Gutiérrez Florido, R., Corpas Pastor, G., and Seghiri, M. (2013). Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain. In *10th Int. Conf. on Terminology and Artificial Intelligence (TIA'13), Workshop on Optimizing Understanding in Multilingual Hospital Encounters*, Paris, France.

Harris, Z. S. (1951). *Methods in Structural Linguistics*. The University of Chicago Press.

Hashemi, H. B., Shakery, A., and Faili, H. (2010). Creating a Persian-English Comparable Corpus. In *2010 Int. Conf. on Multilingual and Multimodal Information Access Evaluation: Cross-language Evaluation Forum*, CLEF'10, pages 27–39. Springer.

Hollmann, W. and Siewierska, A. (2006). Corpora and (the need for) other methods in a study of Lancashire dialect. *Zeitschrift für Anglistik und Amerikanistik*, 1(54):203–216.

Hong, G., Li, C.-H., Zhou, M., and Rim, H.-C. (2010). An Empirical Study on Web Mining of Parallel Data. In *23rd Int. Conf. on Computational Linguistics*, COLING'10, pages 474–482. ACL.

Huang, D., Zhao, L., Li, L., and Yu, H. (2010). Mining Large-scale Comparable Corpora from Chinese-English News Collections. In *23rd Int. Conf. on Computational Linguistics*, COLING'10, pages 472–480, Beijing, China. ACL.

Hunston, S. and Francis, G. (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. John Benjamins Publishing Company.

Hunt, J. W. and McIlroy, M. D. (1976). An Algorithm for Differential File Comparison. Technical Report CSTR 41, Bell Laboratories, Murray Hill, NJ.

Hutchins, W. J. and Somers, H. L. (1992). *An Introduction to Machine Translation*. Academic Press.

Ihalainen, O., Peitsara, K., and Vasko, A.-L. (2006). The Helsinki Corpus of British English Dialects. Reference line and copyright, Department of Modern Languages, University of Helsinki, Helsinki, Finland.

Johansson, S., Atwell, E., Garside, R., and Leech, G. (1986). The Tagged LOB Corpus. Users' Manual. Technical report, Norwegian Computing Centre for the Humanities, Berge, Norway. `http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM`.

Johansson, S., Leech, G., and Goodluck, H. (1978). Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers. Technical report, Department of English, University of Oslo, Oslo, Norway. `http://khnt.hit.uib.no/icame/manuals/lob`.

Johansson, S. and Oksefjell, S. (1998). *Corpora and Cross-linguistic Research: Theory, Method, and Case Studies*. Rodopi.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). The Sketch Engine. In *11^{th} EURALEX Int. Congress*, pages 105–116, Lorient, France.

Kirk, J. M. (1992). *The Northern Ireland Transcribed Corpus of Speech*, pages 65–73. Mouton de Gruyter, Berlin, Germany.

Kjellmer, G. (1986). 'The lesser man': observations on the role of women in modern English writings. In Arts, J. and Meijs, W., editors, *Corpus Linguistics II*, pages 163–176.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit*.

Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD thesis, Publications of the Department of General Linguistics, University of Helsinki, Helsinki, Finland.

Kytö, M. (1996). Manual to the Diachronic Part of The Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts (3^{rd} ed.). Technical report, Department of English, University of Helsinki, Helsinki, Norway. `http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM`.

Lavid López, J. (2005). *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Catédra, Madrid, Spain.

Laviosa, S. (1998). L'Approche Basée sur le corpus/ The Corpus-Based Approach. *Montréal: Les Presses de L'Université de Montréal*, 43(4).

Laviosa, S. (2004). Corpus-based translation studies: Where does it come from? Where is it going? *Corpus-based Translation Studies: Research and Applications*, 35(1):6–27.

Leturia, I., Vicente, I. S., and Saralegi, X. (2009). Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *5^{th} Int. Web as Corpus Workshop*, WAC5, pages 53–61, Donostia/San Sebastian, Spain.

Levenshtein, V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(1):707–710.

Lewis, D. D., Yang, Y., Rose, T., and Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5(1):361–397.

Lüdeling, A. and Kytö, M. (2008). *Corpus linguistics: an international handbook*. Number v. 1 in Handbücher zur Sprach- und Kommunikationswissenschaft. W. de Gruyter.

Ma, X. and Liberman, M. Y. (1999). BITS: A Method for Bilingual Text Search over the Web. In *MachineÂ TranslationÂ Summit VII*.

Maia, B. (2003). What are comparable corpora? In Neumann, S. H.-S. . S., editor, *Workshop on Multilingual Corpora: Linguistic Requirements and Technical Perspectives*, pages 27–34, Lancaster, UK.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. Routledge.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual Terminology Mining - Using Brain, not brawn comparable corpora. In *45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671, Prague, Czech Republic. ACL.

Nie, J.-Y. and Cai, J. (2001). Filtering noisy parallel corpora of web pages. In *IEEE Symposium on Natural Language Processing and Knowledge Engenineering*, pages 453–458.

Nie, J. Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In *22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, SIGIR'99, pages 74–81. ACM.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. Routledge.

Oostdijk, N. and Haan, P. D. (1994). Clause patterns in Modern British English: A corpus-based (quantitative) study. *ICAME Journal*, 18(1):41–79.

Partington, A. (2011). Corpus Linguistics: What it is and what it can do. *Cultus: The Journal of Intercultural Mediation and Communication*, 1(4):35–58.

Pérez Hernández, M. C. (2002). Terminografía basada en corpus: principios teóricos y metodológicos. In Faber, P. and Jiménez, C., editors, *Investigar en terminología*, pages 127–166, Granada, Spain. Comares.

Pirkola, A. (2007). Focused crawling: a means to acquire biological data from the Web. In *Workshop in Data Mining in Bioinformatics*, VLDB'07, Vienna, Austria.

Pirkola, A., Leppänen, E., and Järvelin, K. (2002). The RATF formula (Kwok's formula): exploiting average term frequency in cross-language retrieval. *Information Research*, 7(2).

Pym, A. (2008). *On Toury's Laws of How Translators Translate*, pages 311–328. John Benjamins Publishing Company.

Quirk, R. (1992). On Corpus Principles and Design. In *Directions in Corpus Linguistics*, Nobel Symposium 82, pages 457–469, Stockholm, Sweden. Walter de Gruyter.

Resnik, P. (1998). Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text. In *3rd Conf. of the Association for Machine Translation in the Americas*, AMTA'98, Langhorne, PA, USA. Lecture Notes in Artificial Intelligence 1529.

Resnik, P. (1999). Mining the Web for Bilingual Text. In *37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL'99, pages 527–534. ACL.

Resnik, P. and Smith, N. (2003). The Web As a Parallel Corpus. *Computational Linguistics*, 29(3):349–380.

Rissanen, M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., Nevalainen, T., and Raumolin-Brunberg, H. (1991). The Helsinki Corpus of English Texts. Reference line and copyright, Department of Modern Languages, University of Helsinki, Helsinki, Finland.

Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3):348–379.

Sánchez Martínez, F. and Forcada, M. (2009). Inferring Shallow-transfer Machine Translation Rules from Small Parallel Corpora. *Artificial Intelligence Research*, 34(1):605–635.

Santos, D. and Rocha, P. (2001). Evaluating CETEMPúblico, a free resource for portuguese. In *39th Annual Meeting on Association for Computational Linguistics (ACL'01)*, ACL'01, pages 450–457. ACL.

Saralegi, X., naki San Vicente, I., and Gurrutxaga, A. (2008). Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *1st Workshop on Building and Using Comparable Corpora (BUCC'08)*, pages 27–32, Marrakech, Morocco.

Seghiri, M. (2011). Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. *Revista de Lingüística Teórica y Aplicada (RLA)*, 49(2):13–30.

Sharoff, S. (2010). Analysing similarities and differences between corpora. In *7th Conference of Language Technologies (Jezikovne Tehnologije)*, pages 5–11, Ljubljana, Slovenia.

Silva, J. F. d., Dias, G., Guilloré, S., and Lopes, J. G. P. (1999). Using $LocalMaxs$ Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In *9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, EPIA'99, pages 113–132. Springer.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Skadiņa, I., Aker, A., Giouli, V., Tufiş, D., Gaizauskas, R., Mieriņa, M., and Mastropavlos, N. (2010a). A Collection of Comparable Corpora for Under-resourced Languages. In *2010 Conference on Human Language Technologies – The Baltic Perspective: Proc. 4th Int. Conf. Baltic HLT 2010*, pages 161–168. IOS Press.

Skadiņa, I., Vasiļjevs, A., Skadiņŝ, R., Gaizauskas, R., Tufiş, D., and Gornostay, T. (2010b). Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In *3rd Workshop on Building and Using Comparable Corpora (BUCC'10)*, pages 6–14, Valletta, Malta.

Smith, N. A. (2001). *Detection of translational equivalence.* Undergraduate honors thesis, University of Maryland College Park.

Stamatakis, K., Karkaletsis, V., Paliouras, G., Horlock, J., Grover, C., Curran, J. R., and Dingare, S. (2003). Domain-Specific Web Site Identification: The CROSSMARC Focused Web Crawler. In *2nd Int. Workshop on Web Document Analysis*, WDA'03, Edinburgh, UK.

Summers, D. (2005). Corpus Lexicography - The importance of representativeness in relation to frequency. Pearson education, longman.com/dictionaries, Pearson Education. `http://www.pearsonlongman.com/dictionaries/pdfs/Corpus-lexicography.pdf`.

Svartvik, J., editor (1992). *Directions in corpus linguistics: Proceeding of Nobel symposium.* Mouton Gruyter, Berlin, Germany.

Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., and Keskustalo, H. (2007). Creating and Exploiting a Comparable Corpus in Cross-language Information Retrieval. *ACM Transactions on Information Systems*, 25(1).

Talvensaari, T., Pirkola, A., Järvelin, K., Juhola, M., and Laurikkala, J. (2008). Focused Web Crawling in the Acquisition of Comparable Corpora. *Information Retrieval*, 11(5):427–445.

Taylor, C. (2008). What is corpus linguistics? What the data says? *ICAME Journal*, 1(32):179–200.

Taylor, L. and Knowles, G. (1988). Manual of Information to Accompany the SEC Corpus. The Machine-Readable Corpus of Spoken English. Unit for Computer Research on the English Language, University of Lancaster, Lancaster, UK.

Thomas, J. and Short, M., editors (1996). *Using corpora for language research.* Studies in honour of Geoffrey Leech. Longman.

Torruella, J. and Llisterri, J. (1999). Diseño de corpus textuales y orales. In *Filología e informática: Nuevas tecnologias en los estudios filológicos*, Seminario de Filología e Informática de la Universidad Autónoma de Barcelona y Ed. Milenio, pages 45–77, Barcelona, Spain. José Manuel Blecua, Gloria Clavería, Cárlos Sánchez y Joan Torruella.

Tymoczo, M. (1998). Computerized corpora and the future of translation studies. *Meta: Translators Journal*, 43(4):652–660.

Varantola, K. (2003). *Translators and Disposable Corpora*, pages 55–70. Saint Jerome Publishing.

Wanner, L. (1996). *Lexical Functions in Lexicography and Natural Language Processing*. Companion series. John Benjamins Publishing Company.

Wichmann, A., Fligelstone, S., McEnery, T., and Knowles, G., editors (1997). *Teaching and Language Corpora*. Longman.

Wright, S. and Budin, G. (1997). *Handbook of Terminology Management: Basic aspects of terminology management*. Number V. 1 in Basic Aspects of Terminology Management. John Benjamins Publishing Company.

Wynne, M. (2006). Stylistics: corpus approaches. *Encyclopedia of Language and Linguistics*, 12(2):223–226.

Zanettin, F. (2002). DIY Corpora: The WWW and the Translator. In Maia, B., Haller, J., and Urlrych, M., editors, *Training the Language Services Provider for the New Millennium*, pages 239–248, OPorto, Portugal.

Zanettin, F., Bernardini, S., and Stewart, D. (2003). *Corpora in Translator Education*. Manchester: St. Jerome Publishing.