Alma Mater Studiorum Università di Bologna

SCUOLA DI LINGUE E LETTERATURE, TRADUZIONE E INTERPRETAZIONE

Sede di Forlì

Corso di Laurea magistrale in Traduzione specializzata (classe LM - 94)

TESI DI LAUREA

in Metodi e tecnologie per la traduzione (C.I.)

Bitext alignment: building and evaluating a bilingual corpus and translation memory of academic course descriptions

CANDIDATO

Daniele Cocozza

RELATORE

Prof. Adriano Ferraresi

CORRELATORE

Prof.ssa Silvia Bernardini

Anno Accademico 2013/2014

Sessione III

Contents

Introduction

Following the internationalization of contemporary higher education, academic institutions based in non-English speaking countries are increasingly urged to produce contents in English to address international prospective students and personnel, as well as to increase their attractiveness. The demand for English translations in the institutional academic domain is consequently increasing at a rate exceeding the capacity of the translation profession. Resources for assisting non-native authors/translators in the production of appropriate texts in L2 are therefore required in order to help academic institutions and professionals reduce their translation workload. Some of these resources include: (i) parallel corpora to train machine translation systems and multilingual authoring tools; and (ii) translation memories for computer-aided tools. The purpose of this study is to create and evaluate reference and aiding resources like the ones mentioned in (i) and (ii) through the automatic sentence alignment of a large set of Italian and English as a Lingua Franca (ELF) institutional academic texts given as equivalent but not necessarily parallel (i.e. translated). In this framework, a set of aligners (i.e. tools used to couple pairs of equivalent linguistic entities within a text pair) is examined and compared in order to identify the most profitable one(s) in terms of accuracy and time- and cost-effectiveness. In order to determine the text pairs to align, a sample is selected according to document length similarity (characters) and subsequently evaluated in terms of extent of noisiness/parallelism, alignment accuracy and content leverageability. The results of these analyses serve as the basis for the creation of an aligned parallel corpus of academic course descriptions, which is eventually used to create a translation memory in TMX format.

The present study is part of a larger project funded by the University of Bologna and aimed at creating reference and aiding resources for non-native authors and translators working with English in the institutional academic domain. Chapter 1 reviews the most relevant studies conducted so far on the internationalization and marketization of contemporary higher education (Section 1.1), institutional academic English and ELF (Section 1.2), and translation technology in academic settings (Section 1.3). Detailed information on the present study, its scope and purposes, and the definition of a set of terms and concepts used in this contribution is provided in Section 1.4. Chapter 2 reviews the literature on sentence alignment, examining and

evaluating the set of aligners selected for this study. Full details about the methods used to build and evaluate the parallel corpus/translation memory are outlined in Chapter 3. Results and findings of the analyses conducted on the text pairs and the aligned content are presented in Chapter 4, which also provides detailed information about the creation of the above-mentioned resource(s) and a brief description of their characteristics. Chapter 4 also illustrates the relevance of the findings discussed in this contribution for translation research, the main limitations of this study as well as some suggestions for future research. Finally, the content and findings of this work are summed up in the Conclusion.

Chapter 1 | Background

This chapter will provide the background information for the present study. Building on the internationalization and marketization trends in contemporary higher education (Section 1.1), the basic features of the English institutional academic language will be outlined in Section 1.2. English has indeed become the international lingua franca in academic settings: the international profiling of non-native higher education institutions ultimately depends on its use in the communication with prospective students and personnel. The demand for English translations is therefore considerably high, which makes the task difficult and expensive for most academic institutions worldwide. The automated or assisted translation of institutional academic content into English might be a cost-effective solution in this respect. The effectiveness of these technologies depends on the alignment of a large amount of in-domain data, resulting in parallel corpora or translation memories. Section 1.3 will review the applied studies on machine translation in academic settings and provide a brief description of automated and assisted translation technologies (Section 1.3.1), parallel corpora and sentence alignment (Section 1.3.2). Section 1.4 will provide an overview of the present study, which is part of a larger project, i.e. the CODE project (Section 1.4.1). The scope of the present study, which mostly deals with sentence alignment and translation technology, will be presented in Section 1.4.2. The nature of the text pairs examined and the definition of a set of concepts and terms used throughout this contribution will be provided in Sections 1.4.2 and 1.4.3 respectively. Finally, the contents of this chapter will be briefly summarized in Section 1.5.

1.1 Internationalization and marketization of contemporary higher education

Following the economic, political and societal trends of globalization, higher education institutions are under increasing pressure to develop at the international level. The increasing integration of economies around the world through trade and financial flows has indeed encouraged capital investment in knowledge industries on a global scale. Over the past three decades, supranational organizations and public/private academic institutions worldwide have promoted several internationalization policies and practices to market education beyond national

borders (see Altbach & Knight (2007) for an overview). As a result, the capacity to attract an international audience has become an acknowledged sign of prestige and popularity in higher education.

In Europe, the 1999 inception of the Bologna Process (i.e. the harmonization process of European academic institutions) laid the groundwork for the creation of the European Higher Education Area (EHEA). The EHEA was launched in 2010 as an expression of the accomplishment of a common framework for academic institutions to "ensure more comparable, compatible and coherent systems of higher education [and] to strengthen competitiveness and attractiveness of the European higher education" (EHEA, 2014). Present priorities of the EHEA aim at strengthening and promoting the implementation of mobility strategies among universities as well as increasing the quality of higher education and research EU-wide.[1] Crucially, the achievement of the goals set by this internationalization process ultimately depends on the use of a common language in academic courses. While this is still not the case in most European countries, academic institutions should at least address potential stakeholders using said common language on their websites, which are "a primary source of information for up to 84% of prospective students" (Ferraresi & Bernardini, 2013, and references therein).

Besides internationalization, several scholars have underlined another process undergone by academic institutions since the 1990s, namely the marketization of their discursive patterns. In the early years of this commercial shift in the ethos of higher education, Fairclough (1993) carried out a study relating discourse to the society-driven changes that led British universities to "increasingly […] operate (under government pressure) as if they were ordinary businesses competing to sell their products to consumers" (Fairclough, 1993: 143). By examining the discursive practices of British academic institutions, the scholar notes an historical shift in their social practices. According to Fairclough, this shift is reflected in the entrepreneurial nature/scope of academic institutions. Indeed, "the context of a competitive market where the capacity of a university to attract good applicants is seen as one indicator of its success" has encouraged the use of promotional features in academic discourse (Fairclough, 1993: 156).

[1] Present priorities of the EHEA are part of the 2012-2015 Bologna Follow-Up Group (BFUG) Work Plan. Further details on the Bologna Process, the EHEA, and the BFUG are available at: http://www.ehea.info/ (last visited February 02, 2015).

The marketization trend in the academic discourse has been confirmed a decade later by Swales (2004), who notes that the implicit authority of academic institutions is weakened by the effort to accommodate to the expectations of prospective students and personnel. Likewise, Altbach & Knight (2007: 291) suggest that free trade and international academic mobility led contemporary higher education to be increasingly perceived as "a commodity to be freely traded" (Altbach & Knight, 2007: 291). Similar findings have been reported by other scholars in the first two decades of the 21st century (Mautner, 2005; Morrish & Sauntson, 2013: see Ferraresi & Bernardini, forthcoming, for a brief overview). This suggests once more that the nature and scope of contemporary academic institutions are increasingly defined by what Swales (2004: 8) identifies as the "commodifying trends in higher education" (i.e. the perception of academic institutions as private goods).

The marketization trends in contemporary higher education identified by the critical discourse studies reviewed so far have also found confirmation in several contributions from another branch of research, namely applied corpus linguistics. In a corpus-based study on the hybridization of discursive practices in institutional language and the institutional identity of universities, Caiazzo (2011) suggests that academic institutions are developing a corporate identity through the adoption of communicative strategies from the corporate sector.

1.2 Institutional academic English and English as a Lingua Franca (ELF)

Several social, cultural and political reasons led English to become the predominant global lingua franca in contemporary higher education. As Callahan & Herring (2012: 345) note, "[English] is well established; it confers status and economic advantage; it symbolizes modernity and an international identity; and it is a practical language of cross-cultural communication". More than twenty years after Fairclough's (1993) critical discourse study and following the EU attempt to strengthen student and academic personnel mobility (see Section 1.1), one would expect a substantial increase in the publication of English content on university web pages. A recent study by Callahan & Herring (2012) on the presence (or absence) of multilingual contents on university websites shows promising, albeit still unsatisfactory, results in this respect. On the one hand, the study confirms the

monopoly of English as the international lingua franca of higher education, which is reflected in an overall increase in its use as a primary and secondary language on university websites over a five-year period (i.e. 2006-2011). Specifically, Callahan & Herring (2012: 346) suggest that "[t]he majority of countries where English is not the national language use both their national language(s) and English on their university websites to market to different audiences and for different purposes". On the other hand, it should be noted that the publication of web-based English content does not follow a regular pattern among academic institutions across the world. In this respect the particular case of Europe is worthy of attention. Callahan & Herring (2012) report that the greatest number of academic institutions that use English as a secondary language on their websites comes from North-Western Europe (Norway, Sweden, Denmark, and Finland), surpassing South-Western European countries with an established academic tradition such as Spain, Italy and France. Although the situation might have changed since Callahan & Herring's (2012) study, these results suggest that the internationalization efforts urged by the European Union might not have achieved homogeneous results across Europe. Despite this heterogeneity, some European countries show a certain degree of improvement over the five-year period of the study. For instance, Callahan and Herring's (2012) data reveal that Italy experienced the greatest increase in bilingualism on university web pages.

Despite the progresses made in terms of internationalization by Italian (and, in general, European) universities, further public/private institutional and administrative interventions are required to increase bilingualism and multilingualism in the European academic environment. Against this background, Ferraresi & Bernadini (forthcoming) suggest two lines of research to further the *Englishization* of higher education:

> [o]n the practical/applied side, these [interventions] may include the implementation of tools for assisting non-native writers in producing appropriate texts in this specialized domain [i.e. institutional academic English]; on the descriptive side, studies are required which shed light on the different communicative strategies adopted by universities based in countries where English is used as a native language or as a lingua franca.
>
> (Ferraresi & Bernardini, forthcoming)

Deferring the discussion of the first line of research to Section 1.3.1, it should be noted that at the core of both of them is English institutional academic language. This is defined by Ferraresi & Bernardini (2013) as the language used in

> the wide range of [English] texts used for everyday communication between higher education institutions and their stakeholders, which are likely to feature prominently on university websites – i.e. syllabi, course packs, welcome messages, mission statements, announcements, but also blogs, endorsements, press releases and so forth.
>
> (Ferraresi & Bernardini, 2013: unpaginated)

Research on these genres is nascent, since they have been mainly disregarded over the past three decades. Indeed, most studies conducted so far on the language used in academic settings have investigated expert-to-expert communication and traditional academic genres, such as academic research articles (see Biber (2006) for an overview). Few exceptions are the studies in critical discourse analysis and applied corpus linguistics outlined in Section 1.1, along with other remarkable contributions. Within the latter branch of research, in a pioneering corpus-based study Biber (2006) surveys the distinctive linguistic features of a range of spoken and written institutional academic registers. Specifically, the author provides a functional description of the language used by U.S. institutions to address students, investigating both academic registers (e.g. lectures, textbooks, course reading packets) and institutional registers (e.g. catalogues, syllabi, service encounters) included in the TOEFL 2000 Spoken and Written Academic Language corpus (T2K-SWAL). On the other hand, Afros & Schryer (2009) examine the structural and discursive features of paper-based and web-mediated syllabi of several U.S. universities, concluding that these texts are equally valid resources used by lecturers not only "to manifest their membership in multiple discourse communities, but also to socialize students into (at least, some of) them" (Afros & Schryer, 2009: 231).

Despite investigating less traditional academic genres, the exceptions outlined so far have mainly examined native British and/or North American English. On the other hand, the institutional academic variety of English used by non-native/ELF speakers remains underexplored. Ferraresi & Bernardini (2013) note that several scholars interested in English as a lingua franca (henceforth, ELF) have

explored non-native English varieties in international academic settings. So far, ELF research has examined the institutional academic production from a linguistic perspective. However, it has mainly focused on the spoken medium, neglecting to a large extent written production (but see e.g. Carey (2014)).

An in-depth discussion of the linguistic and pragmatic issues related to ELF lies beyond the scope of the present work (see Jenkins (2011) for an overview). However, a definition of ELF is in order to illustrate the type of language that will be dealt with in the present study.

Seidlhofer (2011: 7) defines ELF as "any use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option". Similarly, the VOICE (Vienna-Oxford International Corpus of English) website[2] offers a definition of ELF as the "English [language] used as a common means of communication among speakers from different first-language backgrounds". ELF is, therefore, "an additional acquired language system" which native English speakers are assumed to use in order to carry out a successful communication in ELF settings (Jenkins, 2011: unpaginated). This implies that

> [a]n alternative view is needed […] that can take account of the ways in which the vast number of ELF users skilfully co-construct English for their own purposes, by treating the language as a shared communicative resource within which they have the freedom to *accommodate* to each other, *code-switch*, and *create innovative forms* that differ from the norms of native English.
>
> (Jenkins, 2011: unpaginated; emphasis added)

On the contrary, the common assumption among native English speakers is that the *Englishization* (and, in general, the internationalization) of academic institutions should take British and/or North American English varieties as models for all universities worldwide (Jenkins, 2011). As Jenkins notes, however, "it is a contradiction for any university anywhere that considers itself *international* to insist on *national* English language norms" (emphasis in the original text).

In order to isolate and describe ELF features in the institutional academic setting, research should therefore examine ELF communication both in isolation and

[2] Further information on the VOICE corpus is available at: https://www.univie.ac.at/voice/ (last visited February 04, 2015).

in relation to native English varieties. To the best of our knowledge, only a handful of studies have been carried out which examine/compare the features of native and lingua franca institutional academic language. These studies belong to a relatively recent and unexplored line of research within applied corpus linguistics. Bernardini et al. (2010) conducted a comparative study on native and non-native English varieties in a set of institutional academic texts produced by British/Irish and Italian universities. The texts were selected and downloaded semi-automatically from the web and included in an annotated monolingual comparable corpus of institutional academic texts (acWaC, i.e. *ac*ademic *W*eb-*a*s-*C*orpus). Drawing on Biber's (2006) analysis of institutional academic registers, Bernardini et al. (2010) compare structural and phraseological patterns and stance expressions in the British/Irish and ELF sub-corpora. The scholars point at the different nature/scope of the language used by native and non-native academic institutions. Findings show indeed that native-English university web pages display more prominently a promotional function and tend to address prospective students/personnel through a personal style. Conversely, ELF texts show a tendency toward a regulatory function, which implies that prospective (international) students of Italian universities are often faced with non-negotiable degree rules and requirements (Bernardini et al., 2010).

In order to extend the analysis of non-native varieties of English used in European academic institutions, Bernardini et al.'s (2010) study was further elaborated on in Ferraresi & Bernardini (2013). Here, the authors present the acWaC-EU corpus – an enhancement of the above-mentioned acWaC corpus that comprises university web pages in English from all European countries. The corpus is used to assess stance expressions in native-English and ELF texts, focusing in particular on modal and semi-modal verbs. Ferraresi & Bernardini (2013) note that obligation/necessity is expressed more directly in ELF texts, which make less frequent use of (semi-)modals on their web pages than Anglophone universities. Moreover, Ferraresi & Bernardini's (2013) evaluation brings to light the heterogeneity of language choices within the ELF discourse community, which may or may not be influenced by the national language(s). Talking about language choices and the use of (semi-)modals by different academic institutions, the authors suggest indeed that "ELF is not a monolithic entity: universities from specific language families may have their own preferences" (Ferraresi & Bernardini, 2013: unpaginated).

In a more recent study, Ferraresi & Bernardini (forthcoming) explore the phraseological patterns in English and ELF university homepages in the acWaC-EU corpus. Findings reveal similar trends to those outlined so far, confirming the existence of differences in communication patterns between native and non-native English. ELF university homepages present indeed a general "underuse of strong collocations [and] overuse of infrequent word combinations", which may be related to linguistic deficiencies of non-native English writers. The analysis also reveals that ELF web-based institutional communication is characterized by an overall use of "more novel combinations, which may or may not be the result of interference from their L1" (Ferraresi & Bernardini, forthcoming).

Bernardini et al.'s (2010) and Ferraresi & Bernardini's (2013; forthcoming) studies confirm Jenkins' (2011) claim that ELF exhibits peculiar features that set it apart from the native standard varieties. However, one of the limits of the studies presented so far is that they adopt a monolingual comparable perspective (i.e. they only explore English texts), leaving out the L1 of ELF texts. Comparative studies are therefore required which investigate the *translation* patterns in non-Anglophone university websites (i.e. national language vs. ELF). The present contribution provides the basic groundwork for a future study of this type (see Section 1.4).

1.3 Translation technology in academic settings

1.3.1 Machine translation and computer-aided tools

The descriptive line of research outlined so far provides useful insights into the language used by non-native English authors in institutional academic communication. These patterns may be identified by comparing the native and lingua-franca versions of university web pages across Europe. As mentioned in Section 1.2, this line of research should be integrated with applied studies paving the way for the implementation of tools to assist non-native authors/translators in the production of appropriate texts in institutional academic English (Ferraresi & Bernardini, forthcoming). A significant example in this respect is the EU-funded Bologna Translation Service (BTS) project, presented in Depraetere et al. (2011).[3] The scholars outline the process leading to the creation of a "web-based, high-

[3] Updated information on BTS is available at: http://www.bologna-translation.eu/ (last visited February 05, 2015).

quality, user-oriented, easily accessible, low cost machine-translation service for the educational domain", starting from the assumption that

> [a]ccess to translated course syllabi and degree programmes plays a crucial role in the degree to which universities effectively attract foreign students and, more importantly, has an impact on international profiling.
>
> (Depraetere et al., 2011: 29-30)

In particular, BTS is meant to be used by universities to translate instantaneously and automatically course descriptions and degree programmes from nine European languages (i.e. Dutch, English, Finnish, French, German, Portuguese, Spanish, Swedish and Turkish) into English. The MT service manages to achieve high-quality results because it "integrates typologically different MT technologies [i.e. statistical and rule-based methods] with translation memory (TM) technology making use of automated post-editing (PE) techniques".[4]

Depraetere et al.'s (2011) contribution is but one of many studies conducted on machine translation (henceforth, MT) from the second half of 20th century to recent years. To the best of our knowledge, however, it is the only study in the natural language processing literature that was conducted on institutional academic language, and more specifically, on course descriptions and degree programmes. The study was motivated by the practical need to provide academic institutions with a tool that would minimize time, cost and effort required for translating a large amount of web-based content into English. In fact, the internationalization of academic institutions has led to a substantial increase in the demand for English translations of web-based institutional academic content (see Section 1.1). Consequently, it is difficult for human translators to produce as many translations as the internationalization trends demand. On the other hand, it might be unprofitable for most non-native academic institutions to pay for human translations, considering the time- and cost-saving potential of MT to streamline their translation workload. Against this background, Depraetere et al. (2011: 31-32) suggest that the automated translation of institutional academic content into English might be beneficial for academic institutions, which would be "in a position to periodically update their

[4] Retrieved from http://bologna-translation.eu/ (last visited February 05, 2015).

existing syllabi and study programmes at a fraction of the time and cost, taking full advantage of leveraging previously produced translations".

A definition of MT might prove useful at this point. In their landmark contribution, Hutchins & Somers (1992: 3) refer to MT as the "computerised systems responsible for the production of translations from one natural language into another". At the core of MT is therefore "the automation of the full translation process" (Hutchins & Somers, 1992: 3). Several MT approaches have been developed over the years (e.g. rule-based, interlingua-based, statistical, hybrid);[5] however, in this study we will mainly refer to statistical MT (see also Section 1.3.2). Several scholars and professional translators acknowledge that MT is only suitable for specific text types and domains. For instance, Kay (prefacing Hutchins & Somers, 1992: xii) claims that MT is meant, among other things, for "material that covers such a narrow subject matter, and in such a routine fashion, as to require little on the part of the translator that could really count as understanding". Likewise, Kohen (2009: 54) stresses that "[h]istorically, many machine translation systems were developed for a limited domain", and as a consequence, "[r]estricting the domain simplifies the machine translation problem dramatically".

Given the complexity of the task, human translators might often find the MT raw output unsatisfactory for publication purposes. In this case, they might want to post-edit (i.e. correct manually) the output if they deem it cost-efficient in terms of time/effort.[6] This evaluation ultimately depends on the quality of the output. Alternatively, they might want to use different aiding resources to increase the productivity and accuracy of their translations, i.e. the so-called translation memory (TM) systems, or more commonly, computer-aided tools (CAT tools). Kohen (2009: 23) defines these tools as "systems that look for matches in large collections of previously translated material". In practical terms, CAT tools act like interactive repositories that help professionals store previous translations[7] and facilitate/speed up the translation process through the automatic retrieval and translation of similar and/or identical instances of previously translated content. As in the case of BTS

[5] See Hutchins & Somers (1992) and Kohen (2009) for an overview.
[6] Several other automatic alternatives can be used to achieve high-quality MT: see Kohen (2009: 27-28) and references therein for an overview.
[7] Previous translations are stored in a *translation memory*, i.e. a collection of source and target paired segments. Among other methods, translation memories can be created through alignment (see Section1.3.2).

(Depraetere et al., 2011), TM technology may also be integrated into MT systems to improve translation automation accuracy.

The restricted domain of the institutional academic genres mentioned in Section 1.2 (i.e. syllabi, course packs, welcome messages, and so forth) leads us to hypothesize that MT systems would produce acceptable results in the institutional academic domain. In addition, since content in most of these texts tends to present only slight variations over different academic years, the statistical retrieval of previously translated material is more likely to return high-quality MT results. However, in line with Bernardini's (2014) idea,[8] MT and similar approaches might still not be the most appropriate solution to provide translated contents on university websites. In fact, the unsatisfactory quality of existing translated texts would affect negatively the quality of the resulting automated translation, and therefore, the international profiling of the academic institution (see also Section 1.4.2).

1.3.2 Parallel corpora and sentence alignment

MT research relies upon large-scale parallel data collection. In fact, a crucial step in the creation of an assisting tool like the one described in Depraetere et al. (2011) (Section 1.3.1) is the collection of a huge amount of in-domain (translated) data. These data take the form of a *parallel corpus*, i.e. "a corpus that contains native language (L1) source texts and their (L2) translations" (McEnery & Hardie, 2012: 20).[9] Among other uses, parallel corpora are used to train MT systems with the ultimate aim of providing substantial information for the statistical retrieval of previously translated chunks and/or entire sentences. Besides their applied (i.e. MT-related) purpose, parallel corpora may also be used by corpus linguists to investigate specific translation patterns from a descriptive perspective. A study of this type, however, lies beyond the purpose of the present contribution.

The creation of parallel corpora and their usefulness for statistical MT systems depend on data alignment (Koehn, 2009), i.e. the (semi-)automatic pairing of equivalent linguistic entities within a text pair. There are as many types of alignment as there are types of segmentation of a text: documents, paragraphs,

[8] See http://code.sslmit.unibo.it/lib/exe/fetch.php?media=code_intro.pdf for further details (last visited February 18, 2015).

[9] The most common form of parallel corpus is unidirectional (i.e. from one language to another). However, corpora can also be bidirectional and multidirectional. See McEnery & Hardie (2012) and references therein for an overview of existing corpora of these types.

sentences, N-grams, syntactic constituents, morphemes or characters (Tiedemann, 2011). In this study we will focus on sentence alignment (see Section 1.4), which we define as follows. Given the sentences s_x, ..., $s_{n(s)}$ (with $x \geq 0$ and $n(s) \geq 1$) in the source text and the sentences t_y, ..., $t_{n(t)}$ (with $y \geq 0$ and $n(t) \geq 1$) in the target text, sentence alignment ($s \parallel t$) consists in a list of paired source and target sentences linked to each other by a relation of translation equivalence and/or uncoupled sentences that have no equivalent in the corresponding source or target text:

$$(s \parallel t) = (s_x, ..., s_{n(s)}) \parallel (t_y, ..., t_{n(t)})$$

According to this definition of sentence alignment, source and target texts should consist of at least 1 sentence each, which may or may not correspond to each other. In this respect, this definition is slightly different from the ones in the literature (e.g. Tiedemann, 2011: 7), since it explicitly accounts for *n*-to-zero and/or zero-to-*n* correspondences. It should be noted indeed that the source and target sentence structure may often differ, i.e. sentences may be split/merged or inserted/omitted during text translation. Consequently, one or more sentences in the source text may be aligned with one or more sentences in the target text. Likewise, one or more sentences in either text may not have a translation equivalent in the other text, and therefore, they may remain uncoupled. The different types of sentence correspondences are defined in Section 1.4.3 below, whereas sentence alignment is extensively covered in Chapter 2.

1.4 Overview of the present study

1.4.1 The CODE project

The present study is part of the CODE (*Cataloghi dell'Offerta Formativa in Europa*) feasibility project[10] funded by the University of Bologna (2013-2015). The CODE project aims at: (i) investigating the phraseology, terminology and semantics of web-based academic course descriptions in national language and ELF in 7 European countries (i.e. Italy, Austria and Germany, France and Belgium, the United Kingdom and Ireland); (ii) establishing quality standards for the ELF production of texts belonging to this institutional academic genre at national and international level (i.e.

[10] Information about the CODE project is available (in Italian) at: http://code.sslmit.unibo.it/ (last visited February 05, 2015).

among universities of the same country and among universities of the countries included in the project); (iii) evaluating said quality standards and cross-verifying their use in the production of other ELF genres. The ultimate aim of the project is to create corpora and tools for assisting non-native professional writers and translators working with institutional academic ELF. At the time of writing, several corpora have been created within the CODE project; in particular:

- **CODE-NAT:** a monolingual corpus of British/Irish web-based academic course descriptions that could be used for the extraction of native standard terminology in the institutional academic domain. The corpus was built partly manually and partly (semi-)automatically. The latter part was derived from both Ferraresi & Bernardini's (2013) acWaC-EU corpus (mentioned in Section 1.2) and the corpus described in Dalan (2012). It includes roughly 11,000 texts from 97 universities (88 from Great Britain and 9 from Ireland) and 1 million tokens, i.e. words (700,000 vs. 300,000 respectively).
- **CODE-ELF:** a monolingual corpus of web-based course descriptions in ELF produced by universities from three linguistic areas, i.e. Italy, German-speaking countries (Austria and Germany), and French-speaking countries (France and Belgium). The corpus consists of 150 course descriptions from each linguistic area, including a series of metadata for each text in the corpus: e.g. subject area (life sciences, physical sciences, social sciences), university country (AU, BE, DE, FR, IT), name of the academic institution, study cycle (first cycle, second cycle, unknown), and text quality defined according to the parameters in the ECTS Users' Guide[11] (low, medium, high). The corpus includes roughly 188,000 tokens (39,000 from German texts; 57,000 from French texts; and 92,000 from Italian texts).
- **CODE-UniBO:** a parallel corpus containing the Italian and ELF versions of the course descriptions published on University of Bologna's website in the 2013/2014 academic year. The corpus includes roughly 4,800 Italian-ELF text pairs and 3.5 million tokens (2 million from Italian texts and 1.5 million from ELF texts).

[11] The ECTS Users' Guide is available at: http://ec.europa.eu/education/tools/docs/ects-guide_en.pdf (last visited February 18, 2015).

1.4.2 Scope of the present study and CODE-UniBO corpus features

Drawing on the resources created within the CODE project, and specifically the CODE-UniBO corpus, the purpose of the present study is to build and evaluate a parallel Italian-ELF corpus of academic course descriptions, which will be created through the automatic alignment of the CODE-UniBO texts at sentence level (see also Chapter 3). In line with the aim of the CODE project, the ultimate aim of this study is to provide a reference resource for CAT tools, MT systems and/or multilingual authoring tools that would support Italian writers in the production/translation of web-based institutional academic content into ELF. Specifically, the purpose of the present study is to contribute to the practical line of research outlined in Section 1.3.1, which is aimed at the creation of resources and the implementation of tools to further the *Englishization* of university websites.

A few considerations on the nature of the text pairs in the CODE-UniBO corpus should be made at this point. First of all, in this study the Italian instances are considered to be the source texts. Indeed, the texts in the CODE-UniBO corpus have been produced by the University of Bologna (i.e. an Italian academic institution). It should be noted that this might not always be the case: we can presume that Italian texts constitute the source texts of the English ones, but we cannot rule out the possibility that English texts are rewritten partially in the target language using source text content only as a reference, or even that they were written in English from scratch. As a consequence, it cannot be assumed that these text pairs are mutual translations (see also Chapter 2). Moreover, English texts might have been produced by either non-native professional translators or university personnel with a more or less advanced knowledge of the English language. Native professional authors/translators might therefore find several language/translation choices questionable. However, the evaluation of the quality of language lies beyond the scope of the present study, where the concept of quality assurance is rather related to the amount of content within the CODE-UniBo corpus that could be stored and leveraged for future use (see also Section 3.2.3). In practical terms, the language used by non-native authors/translators might show a tendency towards replication and/or adaptation of Italian lexical and syntactic constructions, which inevitably results in longer instances than those possibly produced by a native professional. Table 1 shows an example where the target version is longer than necessary and

presents several non-native-like language choices (underlined in the table). In this case, the text includes lexical mistakes (e.g. "the *principle* stages") and syntactic marked constructs (e.g. "*it will be analysed* the evolution"; "its role *on the law development*"; "*the world legal history*").

SOURCE VERSION [IT]	TARGET VERSION [ELF]
La prima fornirà le coordinate storiche dell'esperienza giuridica romana, la nozione di diritto e le sue classificazioni (diritto civile, delle genti, naturale; diritto pubblico e diritto privato), i criteri di buona fede ed equità, le fonti di produzione (sotto il profilo della loro incidenza nel sistema privatistico); delineerà i rapporti fra diritto scritto e diritto consuetudinario, tra "legge" e ordinamento; fornirà la comparazione tra diritto civile romano e diritto delle genti e le notizie fondamentali su le Codificazioni del tardo antico e la somma esperienza della Compilazione di Giustiniano.	The first part will take into consideration the principle stages of Roman legal history from the Law of the Twelve Tablets to the epoch of Justinian. In this context, while not neglecting some considerations on the constitutional changes, it will be analysed the evolution of the concept of law and its classifications (ius naturale , ius civile , ius gentium); the history of the pontifical and secular jurisprudence and its role on the law development; the relation between ius civile and ius honorarium ; the legislative enactments of the emperors in various forms; the vulgar law and post-classical sources; the codes in late antiquity. Special attention will be given to the Justinian age and to the compilation of the Corpus Iuris Civilis , which represents an experience of extraordinary importance for the world legal history.

Table 1. Example of non-native-like language choices in the ELF target text.

Another aspect has to be taken into account in the description of the texts in the CODE-UniBO corpus. Preliminary inspection revealed that it is highly probable to find missing sections/sentences in both the source and target texts (e.g. see uncoupled sentences in Table 2 below) and/or partial translation equivalence between sentence pairs (e.g. Table 2, row 4). Target texts might also be summarized versions of the source texts or present summarized content in the target language (e.g. Table 2, row 20). What is more, the structure and content of source texts may not be reflected in the target text, leading to cross-links, re-orderings and overlapping information, among other things, which may ultimately affect the alignment quality. Table 2 shows, by way of example, a manually aligned text pair and (some of) the features outlines so far: the progressive numbers in the first column are reported for references purposes, the (presumed) Italian source text and

(presumed) EFL target text are presented in the second and third column respectively.

	SOURCE TEXT [IT]	TARGET TEXT [ELF]
1	###Corso di laurea: PROGETTAZIONE E GESTIONE DELL'INTERVENTO EDUCATIVO NEL DISAGIO SOCIALE ###Titolo: PROCESSI COGNITIVI DISFUNZIONALI	###Corso di laurea: PROGETTAZIONE E GESTIONE DELL'INTERVENTO EDUCATIVO NEL DISAGIO SOCIALE ###Titolo: PROCESSI COGNITIVI DISFUNZIONALI
2	###Programma	###Programma
3		The course surveys many fundamental areas within the field of cognitive processes.
4	Saranno affrontati i pricipali modelli teorici della psicologia dell'apprendimento, dal comportamentismo al modularismo di impostazione cognitivista. Saranno quindi analizzati i principali disturbi dell'apprendimento e le tecniche di valutazione e di intervento educativo.	A principal focus is on how cognitive psychology explains specific learning deseases and on which strategies normally are used for the rehabilitation of specific cognitive functions.
5	In specifico saranno trattati i disturbi specifici di apprendimento a base genetica, biologica o ambientale.	
6	Disturbi del linguaggio, della letto-scrittura, discalculia e difficoltà nel problem solving e nella comprensione, deficit di attenzione, iperattività e alcune sindromi genetiche che hanno ripercussioni sull'apprendimento.	Dyslessia, Development Dyscalculia,Attention Deficit Hyperactivity Disorder - ADHD, language disorders, problem solving and comprehension difficulties will be considered.
7	###Metodi	###Metodi
8	###Tipo	###Tipo
9	Per la verifica è prevista una prova orale	
10	Per gli studenti non frequentanti all'esame orale si aggiunge una Relazione scritta su un disturbo di apprendimento da consegnare una settimana prima dell'esame, la relazione deve seguire lo schema seguente:	
11	1- Manifestazione comportamentale (poche righe)	
12	2- Cosa non funziona? Processi cognitivi compromessi	
13	3- Strumenti di intervento	
14	Lunghezza: complessivamente 2 cartelle	
15	###Obiettivi	###Obiettivi
16		The aim of course is to offer an analytic and up-to-date survey of knowledge in the field of learning and memory deseases.
17	Al termine del corso lo studente:	At the end of the course participants will learn:
18	- conosce i principali paradigmi psicologici relativi ai processi di apprendimento e memoria e ai	

	loro deficit;	
19	- è in grado di progettare interventi educativi e formativi nell'ambito dei disturbi di apprendimento;	a) to project an educational plan for the learning deficit
20	- è in grado di valutare progetti di intervento riabilitativo nell'ambito dei disturbi cognitivi e in particolare in quelli relativi all'apprendimento;	b) to evaluate a rehabilitation program
21	- è in grado di utilizzare strumenti di analisi e comparazione per approfondire in autonomia le proprie conoscenze nell'ambito del disagio prodotto da disturbi della funzione cognitiva.	
22		c) to monitor the efficacy of educational plan for learning difficulties depending on biological and/or environmental deseases
23	###Supporti	###Supporti
24	Saranno utilizzati programmi power point che gli studenti potranno consultare nel sito della facoltà come supporti alla didattica	This course will include lectures, and PowerPoint presentations

Table 2. Example of a manually aligned text pair in the CODE-UniBO corpus.

Table 2 also gives a general idea of the structure and content of the extracted text pairs. First of all, texts in the CODE-UniBO corpus often present missing punctuation (e.g. rows 9 and 11-14). This might be problematic in terms of automatic alignment. Indeed, the aligning algorithm/software might not recognize the end of the sentence, and consequently, misalign the files (see also Section 3.2.2). On the other hand, these texts may include bulleted, numbered or simple list elements (e.g. rows 11-14 and 18-22). It should be also noted that *all* the texts include several metadata (preceded by three hash signs "#" in Table 2). Information on the name of the course and the title of the degree programme is part of row 1. Metadata also mark the different sections imposed by the UniBO template of course unit descriptions: course contents (row 2), methods (row 7), type of exam (row 8), aims (row 15), and teaching supports (row 23). By way of example, the Italian course description presented in Table 2 is shown in Figure 1 as it appears on the UniBO website.

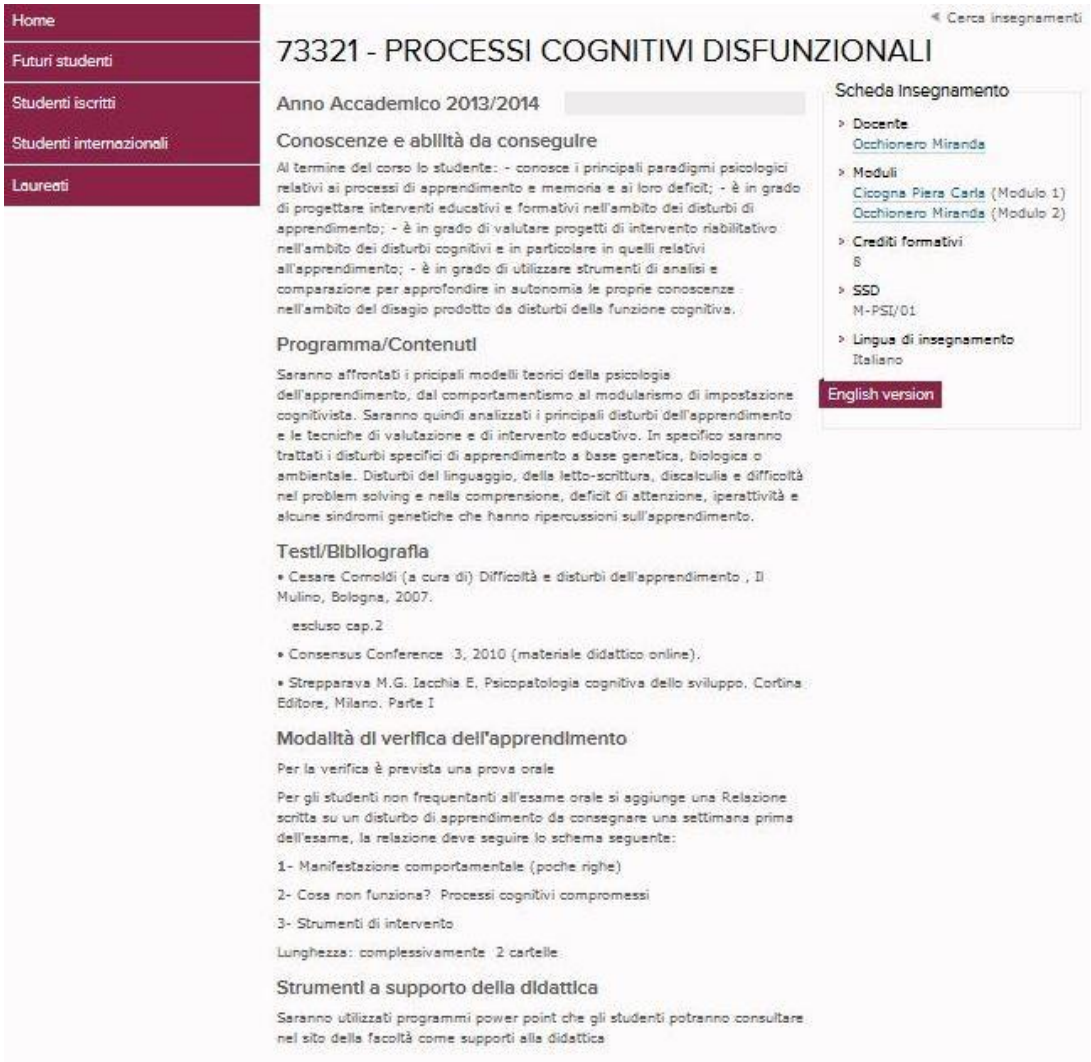

Figure 1. Example of Italian course description as it appears on the UniBO website.

As can be noticed, the texts extracted do not contain all the sections imposed by the UniBO template of course unit descriptions (e.g. the bibliography section "Testi/Bibliografia" is not included). This was a choice that the research team that built the CODE-UniBO corpus took to reduce noise. Nevertheless, it might still be possible to find items in bibliographic reference lists in the other sections imposed by the template and included in the corpus. These items are considered to be sentences in the present study together with metadata and headings/sub-headings (see Table 3 below). Section 3.2.1 presents further information and a definition of the notion of sentence as regarded in the present study.

Metadata	###Corso di laurea: PROGETTAZIONE E GESTIONE DELL'INTERVENTO EDUCATIVO NEL DISAGIO SOCIALE ###Titolo: PROCESSI COGNITIVI DISFUNZIONALI	(*1 sentence*)
Heading	Part A (30 hours)	(*1 sentence*)
Bibliographic reference	B. Zimmermann, La commedia greca. Dalle origini all'età ellenistica , Roma (Carocci) 2010,	(*1 sentence*)

Table 3. Examples of metadata, heading and items in bibliographic reference lists in the CODE-UniBO corpus.

Last but not least, these texts might present several character encoding errors such as those underlined in the Italian sentence in Table 4.

CHARACTER ENCODING ERRORS
Il corso si propone di fornire agli studenti gli strumenti per la conoscenza, l?analisi e la lettura critica delle architetture del passato, dall?Antico fino all?età barocca, con particolare attenzione alle loro parti costituenti, formali e strutturali; ai materiali e alle tecniche costruttive impiegate; agli obiettivi del committente, del costruttore e dell?architetto; alle relazioni con le architetture precedenti e coeve.
- Strutture e processi organizzativi della funzione R&S
- Simultaneous & concurrent engineering: aspetti organizzativi e tecniche operative

Table 4. Example of character encoding error in an Italian text.

The correct alignment of these sentences ultimately depends on the fine-tuning of the aligner's segmentation rules (see Section 3.2.2).

1.4.3 Definitions and concepts

It should be clear by now that the present study deals with bitext alignment (sentence) and translation technology. The terminology used for each branch of research is briefly outlined in what follows.

Alignment terminology is mostly based on Tiedemann (2011) and adapted to the specific needs of the present project. Specifically, the following terms may be found when talking about alignment:

▪ **bitext**	a bilingual pair of texts that presents full or partial translation equivalence and composed of different types of segment/sentence correspondences. Tiedemann's (2011: 7) "symmetric relation" between source and target texts is not a prerequisite of the present study (i.e. texts often present missing/extra sections or sentences). On the contrary, we adopt the following assumption

	from Tiedemann (2011: 7): "we usually do not require that one half of the bitext is the original source text and the other half is the target text that has been produced on the basis of that source text. However, it is often convenient to think of source and target texts when talking about bitexts" (see also Section 1.4.2). Bitexts may also be referred to as **text pairs**.
▪ **bitext half**	the source and/or target text within a bitext (Tiedemann, 2011).
▪ **segmentation**	"[the] division of text into meaningful units" (Tiedemann, 2011: 23). In this study segmentation will be performed at sentence level (see Section 3.2.2).
▪ **alignment**	"[t]he entire structure that connects both bitext halves with each other according to some notion of correspondence" (Tiedemann, 2011: 24). The term does not refer to individual items linked together, but to the whole set of linked items (Tiedemann, 2011: 7).
▪ **lexical cues**	pairs of equivalent lexical items that are used to identify corresponding source and target segments/sentences within the bitext (Tiedemann, 2011).
▪ **bitext link**	aligned pair of source and target items, i.e. segments/sentences. Bitext links may also be referred to as **segment links** and **sentence links** respectively. The term accounts for any type of bitext link correspondence.
▪ **bitext link correspondence** or ▪ **segment/sentence correspondence**	- **one-to-one correspondence:** one item in the source text is linked with exactly one item in the target text; - **one-to-many** or **many-to-one correspondence:** one item in the source text is linked with two or more items in the target text, and vice versa; - ***n*-to-zero** or **zero-to-*n* correspondence (where *n* ≥ 1):** one or more items in one bitext half have no correspondence in the other bitext half; - **many-to-many** or ***n*-to-*n* correspondence (where *n* ≥**

	2): more than one item in one bitext half is coupled with more than one item in the other bitext half.
▪ **cross-link**	a bitext link that contains inverted items within the bitext (e.g. a source item in position (1) is coupled with a target item in position (2)).

Some of the previous terms may be replaced when talking about translation technology, including alignment technology. Hence, where relevant, *bitexts* may also be referred to as **input files** or **file pairs** (i.e. any Italian-ELF file pair of machine-readable data imported into an alignment tool). Accordingly, *bitext halves* may be also referred to as **source** and **target (input) files**. Finally, when talking about translation memories and CAT tools, the term *bitext link* will be replaced by the term **translation unit** (i.e. an aligned pair of source and target segments that may include any type of sentence correspondence).

1.5 Summary

In this chapter, we have seen how the internationalization of higher education (Section 1.1) and the monopoly of English as the global lingua franca (Section 1.2) have led to a substantial increase in the demand for English translations in the institutional academic domain. As a consequence, universities might find difficult and/or unprofitable to *Englishize* their websites through human translation. Hence, several translation technology solutions have been presented (i.e. MT systems, CAT tools) which may help them streamline their translation workload (Section 1.3). These systems rely on large amounts of domain-restricted data, which take the form of parallel corpora and/or translation memories created through (semi-)automatic alignment. As part of a larger project that aims at providing reference and aiding resources for non-native authors and translators in the English institutional academic domain (i.e. the CODE project), the present study aims at building and evaluating a parallel corpus/translation memory, created through the automatic alignment of a large set of Italian-ELF web-based academic course descriptions (Section 1.4). Given the noisy nature of these items, the features of a set of sentence aligners have been examined in order to detect the tool(s) that would give the best accuracy results with the text pairs under consideration. The analysis and comparison of these aligners will be presented in Chapter 2.

Chapter 2 | Bitext sentence alignment: analysis and evaluation of a set of aligners

In this chapter, several alignment tools will be presented and compared to each other with the aim of selecting the one that best meets the needs of the present project, presented in what follows.

Preliminary inspection revealed that parallel data in the CODE-UniBO corpus are often noisy (see Section 1.4.2). Thus, we decided to extract automatically all the corresponding pairs of Italian and ELF texts having roughly similar length in terms of characters, and specifically a maximum delta (i.e. difference) of ±40 per cent. This was based on the hypothesis that mutual translations would have roughly the same number of characters. The bitext extraction from the university database (i.e. the CODE-UniBO corpus) eventually resulted in 3,263 alignable items. Due to the large number of bitexts, the manual revision/correction of their automatic alignment was deemed unprofitable in terms of human effort. A prior analysis of the pros and cons of a set of aligner was therefore crucial to detect the tool(s) that would yield the highest alignment accuracy. In fact, although the 40-percent-delta parameter gives us text pairs which are highly likely to be translations of each other (based on document length similarity), sentence-level translation equivalence cannot be taken for granted. The main challenge of our work is therefore to find a method for identifying and aligning parallel sentences in noisy corpora: our aim is to automate as much as possible the alignment process, maximizing accuracy with the minimum amount of human effort. Bearing this in mind, a list of alignment tools was compiled on the assumption that the selected tools operate (at least) at sentence and/or coarser granularity level. In several occasions it was found that the aligner also operated at sub-sentence level, but this did not prevent us from including it in the list. However, no specific tree-level, phrase-level and word-level aligner was considered or included in the set of aligners to be analyzed for our purpose. In the case of commercial solutions, we imposed the further constraint that the tool(s) be available on the Department's PCs and/or their price be reasonable for a feasibility research study. Likewise, many commercial products were discarded as expensive or unavailable.

The resulting list of aligners thus includes traditional and more recent sentence alignment algorithms and standalone applications that exist at the time of

writing. In order to examine these aligners, a set of user-oriented parameters is defined in Section 2.1. The analysis and evaluation of the set of aligners follows in Section 2.2 and Section 2.3 respectively. The list is also available in table format on the left-hand side of Appendix A.

2.1 Parameter definition

As already mentioned, the definition of a set of parameters is in order to compare the features and performance of each aligner. During this phase, it was decided that all the aligners had to ensure a degree of automation of the process (given the nature of the bitexts and the frequent lack of parallel data, a certain degree of human effort is still often required). Table 1 shows a user-oriented categorization of the criteria selected for the comparison and evaluation of the aligners: the parameters are grouped in three categories (software characteristics, basic features, and advanced features) and a definition is provided along with each parameter.

SOFTWARE CHARACTERISTICS	
Approach	this parameter defines the specific technique and cues used by the aligner to carry out the alignment task (namely sentence length, lexical information or a combination of both), as well as other criteria such as structural and formatting information, punctuation, and so forth (see Section 2.2 for a brief description of each type of approach).
Programming language and requirements	the language in which the aligner is written and/or specific software requirements. The programming language is not further commented on in the analysis presented in Section 2.2, but it is reported in Appendix A for the sake of thoroughness.
Interface	the parameter distinguishes between applications that support a command-line user interface (CLI) and those that use a graphical user interface (GUI). In the case of GUIs, it also specifies whether or not source and target bitext links can be edited interactively.

Availability	it distinguishes free tools from commercial products.
BASIC FEATURES	
Granularity level	this parameter specifies the linguistic entity according to which the aligning algorithm or standalone application segments the bitext halves and couples source and target text elements. As mentioned in the introductory section of Chapter 2, a common feature of the selected aligners is that they operate at sentence level, but they may also handle coarser or more fine-grained granularity levels.
Language pairs	it defines whether the aligner is language independent or not. In the latter case, it presents the supported language pairs.
File pairs per alignment	it shows the number of bitexts that can be aligned in each alignment session.
Bitext link correspondence	it lists all the supported text linking relations in the alignment raw output (see Section 1.4.3). This parameter also indicates, where relevant, the features that enable users to solve the problem of cross-linking.
Output format(s)	it defines the available output formats. Given the nature of our task, particular attention is paid to the availability of TMX export options, as well as the possibility to merge several output documents into a single file.
ADVANCED FEATURES	
Additional resources	any required and/or recommended built-in or external resource on which the system relies to perform the alignment or to improve its accuracy.
Pre-processing	it lists any required or recommended corpus pre-segmentation and/or pre-processing operation on which the system relies to perform the alignment or to improve its accuracy.
Segmentation rules	if available, it presents the chance to set up

configuration	segmentation options for better alignment quality.

Table 1. User-oriented parameters for the comparison of the features and performance of the selected aligners.

As mentioned above, these parameters will serve as the basis for the analysis and comparison of the selected aligners discussed in Sections 2.2 and 2.3. Notice, however, that our study is not focused on an extensive comparative investigation of each alignment system and that our purpose is not to provide an in-depth description of the nature and performance of their algorithms and characteristics. Rather, our contribution aims to provide a functional comparison of the features and performance of the aligners as they are reported in the literature. Such comparison is presented in Section 2.2 and, in summary form, in Appendix A. Specifically, Section 2.2.1 presents the features and performance of the aligning algorithms and CLI aligners from our list; Section 2.2.2 describes a set of free GUI alignment tools; and Section 2.2.3 investigates the features of a series of commercial GUI alignment tools. Finally, the suitability of the whole set of aligners for the present project will be discussed in Section 2.3.

2.2 Features and performance of the aligners

2.2.1 Aligning algorithms and CLI aligners

Gale & Church (1991) achieved a major breakthrough in computational linguistics by introducing a statistical length-based alignment technique. In their words,

> [t]he model makes use of the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to each pair of proposed sentence pairs, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences.
>
> (Gale & Church, 1991: 178)

The fact that sentence length (measured in characters) is the only feature used by the algorithm (see also Tiedemann, 2011: 39) makes the algorithm itself context and language independent. One of the weaknesses of such technique is, however, the risk of wrong bitext mapping in cases where the aligner is faced with neighbouring sentences containing roughly the same number of characters. In such cases, the length ratio does not provide enough information for the algorithm to make the correct alignment decisions. This results in a high risk of misalignment (Wu, 2010; Tiedemann, 2011).

Sentence-length correlation between bitext halves was also at the basis of a concurrent study conducted by Brown et al. (1991) and presented alongside Gale & Church's (1991) work at the 29th Annual Meeting of the Association for Computational Linguistics. Unlike the latter approach, Brown et al. (1991) rely on words to determine length-base bitext link correspondences. In a subsequent study by Gale & Church (1993), such method was found to be less accurate than the character-based one, and it was proved that character-based length detection is probabilistically more robust given that the number of characters in a sentence is larger than the number of words. Nonetheless, as Wu (2010: 382) observes, "[their] experiments were conducted only on English, French, and German, whose large number of cognates improve the character length correlation". For this reason, Gale & Church's algorithm is more likely to yield higher accuracy with Indo-European languages. It should be further noted that Gale & Church (1991) tested their algorithm on texts that show a high degree of translation equivalence, i.e. the Canadian Hansards parliamentary proceedings. Broadly speaking, length-based alignment methods perform well assuming that both bitext halves present the same structure and amount of information. In particular, Gale & Church's algorithm relies upon two kinds of constraints, namely the *bijectivity* and the *monotonicity* constraint. According to the former, the system maps bitexts on a one-to-one segment correspondence, falling therefore into the category of bisegmentation (Tiedemann 2011). On the other hand, the monotonicity constraint assumes that both bitexts halves present corresponding segments in the same order, which excludes any cross-translation, insertion and omission by the translator(s). Here lies possibly the strongest weakness of length-based alignment methods: when faced with noisy bitexts, the misalignment rate increases, translating into a high risk of error propagation.

A possible solution to improve the quality of the alignment consists in using coarse-grained synchronization points (i.e. paragraph boundaries) to guide iteratively the algorithm through the alignment of more fine-grained linguistic entities (i.e. sentences) (Tiedemann, 2011). Alternatively, the alignment quality could be enhanced by combining length and lexical cues in what Wu (2010) calls "multifeature sentence alignment". Broadly speaking, there are two possible methods that implement the latter approach. The first one can be considered as an enhancement of simple length sentence alignment inasmuch as it first couples sentences through a length-based algorithm, and then it uses lexical cues in the form of words or cognates to improve the results of the initial alignment (Wu, 2010). An example of this approach is at the core of **Moore's bilingual sentence aligner** (Moore, 2002). The latter algorithm relies on a modified version of Brown et al.'s (1991) length-based aligner to find high-probability one-to-one correspondences that are used in a second stage to train a word-translation model and to find lexical correspondences between the bitext halves. Combined with sentence length information, such correspondences are subsequently used to realign the corpus (Moore, 2002). Since word correspondences are generated automatically as a by-product, no additional lexical resource is required and the method can be assumed to be language independent. Moore's bilingual sentence aligner requires corpus pre-tokenization and pre-segmentation so that each line contains a sentence and space delimits word boundaries. Finally, as in the case of Gale & Church's algorithm, Moore's sentence aligner operates on the monotonicity constraint, i.e. it assumes that the sentences to be aligned are in the same order in both bitext halves.

Wu (2010) identifies another method for multifeature sentence alignment, which relies upon the integration of length and lexical information. In this case, an algorithm identifies lexical candidates through an external or automatically generated bilingual lexicon and uses them iteratively to find a series of alignment reference points in the bitexts (Wu, 2010). As Tiedemann (2011) explains, in a subsequent stage the resulting bitext lexical map is used to align bitexts at sentence level. This can be performed through sentence length information or geometric bitext mapping of lexical points of correspondence (Melamed, 1996). The former approach is at the core of **Champollion Tool Kit (CTK)** and **Translation Corpus Aligner (TCA)**, whereas an example of the latter method is implemented in the

Geometric Mapping and Alignment (GMA) approach. Each of these methods will be described in greater detail in what follows.

CTK is built around the Champollion sentence aligner algorithm (Ma, 2006). It relies mainly upon lexical cues, even though it also uses sentence length information to optimize alignment results. An external or built-in dictionary is required, which implies that the approach is language dependent. Unlike the algorithms examined so far, a positive aspect of the tool kit is the fact that it assumes noisy parallel content featuring sentence insertions and deletions. It outputs one-to-zero/zero-to-one, one-to-one and two-to-one/one-to-two bitext link correspondences. Beside the translation lexicon, the tool kit requires input files to be sentence-segmented with newlines at the end of each segment. Tokenization (i.e. word segmentation) is not required, but it could help improve both precision and recall. To the best of our knowledge, CTK currently supports three language pairs, namely English-Chinese, English-Arabic, and English-Hindi, even though additional language pairs can be added.

TCA relies upon lexical information from a bilingual lexicon (referred to as "anchor list") together with length information and other parameters such as cognates, capitalization, punctuation, and structural and formatting information (Hofland & Johansson, 1998). Since there is no direct or visible link to the download page of the aligner, we consider it to be currently unavailable. However, it is worth presenting its features in this overview. Similarly to CTK, TCA is language dependent. The aligner was originally intended for the English-Norwegian language pair, but it has since been extended to other combinations with English as a common language, including German, Dutch, Portuguese, and Spanish. Language dependency, however, is not the only feature that TCA has in common with CTK. Indeed, both aligners require sentence pre-segmentation, even though TCA also needs input files to be marked up in XML format. Unlike CTK, TCA allows multiple file alignment and text link interactive editing. The system only assumes one-to-one correspondences to be correct, whereas one-to-two and one-to-zero correspondences have to be manually checked. These are the only instances that the system includes in the file to be reviewed (Santos & Oksefjell, 1999).

As mentioned above, a different approach underlies Melamed's (1996) **GMA**. The method implements two algorithms to perform the alignment. In the initial stage, the system uses the so-called Smooth Injective Map Recognizer (SIMR)

algorithm to create a series of possible lexical connections between the two bitext halves (based on high translation equivalence probability). These connections are used to select a set of true points of correspondence (TPCs) through which the system creates a map of the bitext. In the second stage, the Geometric Sentence Alignment (GSA) algorithm performs the bitext alignment at sentence level on the basis of the information contained in the lexical map created in the initial stage (Melamed, 1996). As Melamed (1996) remarks, a perfect sentence alignment can ideally be obtained through a complete set of TPCs combined with suitable boundary information. However, since SIMR often finds incomplete and noisy correspondence points, its output presents a series of alignment errors. When this is the case, GSA re-aligns misaligned bitext links through Gale & Church's (1991) length-based algorithm. Melamed (1996), however, openly criticizes sentence alignment as being "of dubious practical value", arguing that "a set of correspondence points, supplemented with sentence boundary information, expresses sentence correspondence, which is a richer representation than sentence alignment" (Melamed, 1996: 8). Sentence correspondence is indeed the result of SIMR's bitext mapping, which, unlike GSA's sentence alignment, also accounts for cross-links (Melamed, 1996). The aligner has been tested on several language pairs with English as a common language (see Appendix A). No corpus pre-processing or additional resource is required, although a translation lexicon and a list of stop words is recommended for better results.

A different, hybrid approach is at the core of Varga et al.'s (2005) aligning algorithm **hunalign**.[12] Depending on the availability of a dictionary, hunalign uses two distinct approaches. If a dictionary is added to the sentence aligner, the algorithm combines the related lexical information and Gale & Church's (1991) length-based approach to perform the alignment task. Conversely, if a dictionary is not available, hunalign first makes use of sentence-length information, and subsequently bootstraps a dictionary based on the results of the initial alignment. In a second pass, the bitext is re-aligned using the bilingual information contained in the bootstrapped dictionary (Varga et al., 2005). Hunalign uses a similar approach to Moore's (2002) bilingual sentence aligner insofar as they both combine length and lexical cues. However, as Varga et al. (2005) maintain,

[12] Updated information on hunalign is available at: http://mokk.bme.hu/resources/hunalign/ (last visited February 22, 2015).

> [Moore's] simpler method using dictionary-based crude translation model [...] has the very important advantage that it can exploit a bilingual lexicon, if one is available, and tune it according to frequencies in the target corpus or even enhance it with extra local dictionary bootstrapped from an initial phase.

Hunalign is basically language independent and its implementation only requires space-delimited pre-tokenization at word level and newline-delimited pre-segmentation at sentence level. Its outstanding feature is the ability to recognize both *n*-to-zero/zero-to-*n* and many-to-one/one-to-many translation equivalence between bitext links. Here lies another important difference with Moore's sentence aligner: in Varga et al.'s (2005) terms, "the focus of Moore's algorithm on one-to-one alignment is less than optimal, since excluding one-to-many and many-to-one alignments may result in losing substantial amounts of aligned material if the two languages have different structuring conventions". Despite hunalign's high accuracy in terms of both precision and recall (i.e. 97% to 99%; see Varga et al., 2005), it might be sometimes necessary to modify and/or improve alignment results, especially when the algorithm is faced with noisy corpora. Last but not least, it is worth noting that hunalign requires an aiding tool (**partialAlign**) when faced with input files containing more than ten thousand unaligned sentences (MOKK, 2015).

Similarly to GMA (Melamed, 1996), geometric information is also at the core of **Tagaligner**.[13] This application uses XML-based tag structure and text block length to improve sentence-level alignment. Unfortunately, not much information is known about this tool, except for the fact that it creates sentence-aligned TMX files starting from markup-language-based input files – mostly in XHML and HTML format.

Little is known also of the bilingual sentence-alignment systems **Align**, **Gargantua**, and **bligner**. At the core of Align is a user-defined scoring function that guides the alignment on a sentence-by-sentence basis (Berger, 2000). The system assumes omissions in the bitext and attempts to align segments containing high-probability bilingual word-to-word translation equivalence based on said scoring function (Berger, 2000). Similarly to other aligners, input files need to be sentence-

[13] Information on TagAligner is available at: http://tag-aligner.sourceforge.net/ (last visited February 22, 2015).

segmented. No information is available on the supported language pairs, the bitext link correspondences and the available output format(s).

On the other hand, **Gargantua** integrates a modified version of Moore's (2002) method in a two-step approach to the alignment of one-to-zero/zero-to-one, one-to-one, and many-to-one/one-to-many sentence correspondences (Braune & Fraser, 2010). The aligner is intended for symmetrical and asymmetrical corpora and requires sentence-segmented input files with one sentence per line and space-delimited words (a pre-processing operation that requires little effort from users that can handle regular expressions).

As for **bligner**, the system generates TMX files through the alignment of bitexts at sentence or paragraph level.[14] Being a batch aligner and operating through command lines, the output bitext links cannot be interactively edited. However, segmentation rules can be configured, which might increase the alignment quality.

2.2.2 Free GUI alignment tools

Due to hunalign's (Varga et al., 2005) high alignment output quality (see Section 2.2.1), several developers have bundled its algorithm into their own applications. It is the case of András Farkas' **LF Aligner** (MOKK, 2015), Vondřička's (2014a; 2014b) **InterText**,[15] and **eAlign**.[16] As one might expect, the alignments performed by the three applications ultimately depend on hunalign's features and output (see Section 2.2.1). Unlike hunalign, however, no additional resource (i.e. partialAlign) is required to align a large set of text pairs. Moreover, since these wrappers are designed as interactive tools, users can manually correct misaligned bitext links to improve alignment accuracy. The features of each of these aligners will be briefly outlined in what follows (see also hunalign's description in Section 2.2.1 and data in Appendix A).

LF Aligner incorporates both hunalign and partialAlign (MOKK, 2015), it can align texts in up to 100 languages simultaneously and it includes built-in dictionary data for more than 800 combinations of 32 languages. The system also

[14] Information on bligner is available at: http://www.bligner.org/ (last visited February 22, 2015).
[15] In the present study we refer to the **InterText Editor** version. Detailed information on the different versions of InterText is available at: http://wanthalf.saga.cz/intertext (Vondřička, 2014b; last visited February 22, 2015).
[16] Formerly known as **SuperAlign**. Information available at: http://sourceforge.net/projects/ealign/ (last visited February 22, 2015).

deals with document format conversion and customizable sentence segmentation options.[17]

On the other hand, InterText and eAlign can be said to be roughly equivalent in terms of the nature and performance of the integrated features. InterText is an alignment editor developed to manage the alignment of bitexts (including multiple parallel language versions of the same text) at sentence and/or paragraph level. The software integrates a fully configurable sentence splitter based on complex regular expressions (see Figure 1) that might help improve alignment accuracy.

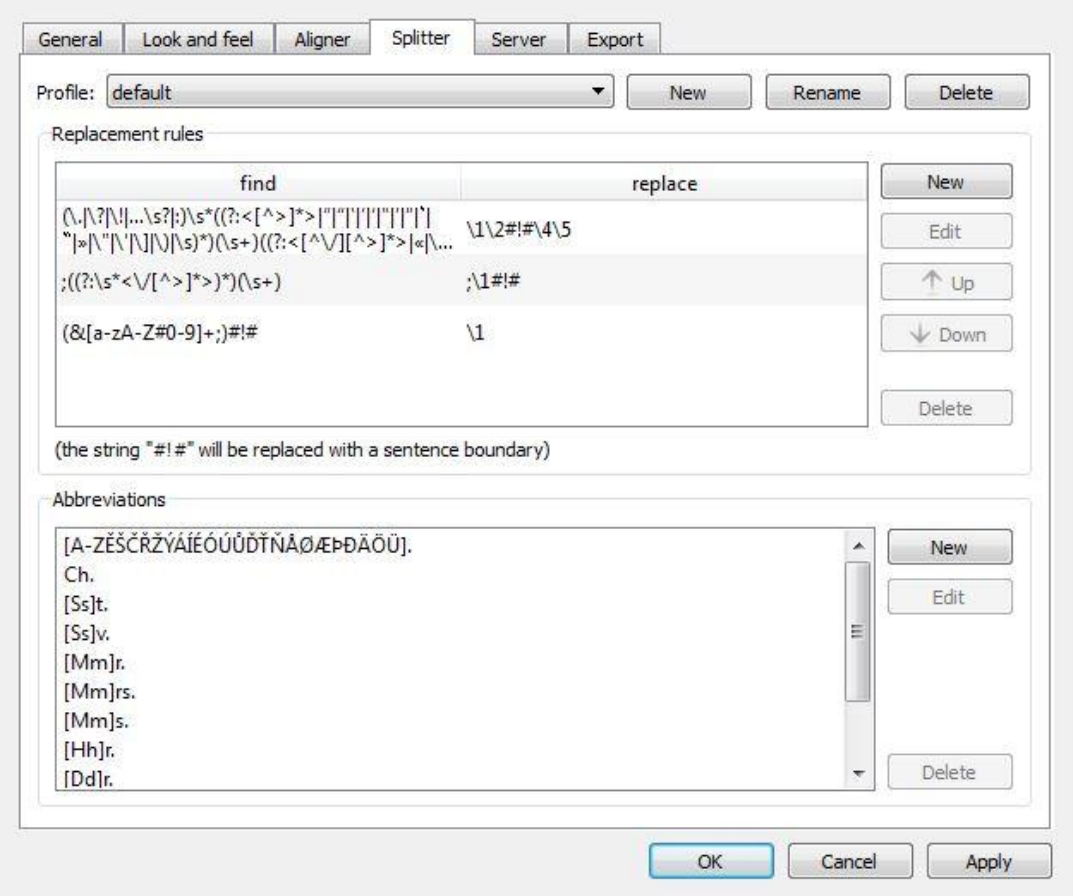

Figure 1. InterText's splitter settings.

Likewise, eAlign is an interactive tool developed to perform sentence/paragraph-level alignments of parallel corpora. Users can easily specify the punctuation marks to be considered as sentence breaks and list a series of abbreviations and exceptions that the systems would skip during segmentation (see Section 3.2.2 for further details). Although both applications perform one alignment at a time, they can handle multiple files, which speeds up the bitext import process

[17] Information retrieved from http://sourceforge.net/projects/aligner/ (last visited February 22, 2015).

and facilitates the management of multiple alignment tasks. Both aligners export alignments to TMX format, among other file formats (see Appendix A). Unlike InterText, however, eAlign also allows output alignments to be merged into a single TMX file.

Little or nothing is known about the approach at the core of two different alignment tools, namely **Align Assist** (GITS, 2008) and **bitext2tmx** (Forcada & Martin, 2010). Bitext2tmx aligns two plain text files at sentence level and generates a translation memory in TMX format. A quick test suggested that this tool seemingly uses punctuation as the main segmentation option. Bitexts are indeed poorly segmented if punctuation is missing, as in the case of the CODE-UniBO corpus (cf. Section 1.4.2), and a high degree of human effort is required to correct misalignments. Based on our experience, however, alignment quality slightly improves with pre-processed input files where line breaks delimit sentence boundaries.

Similarly to bitext2tmx, Align Assist creates translation memories (in TMX format) through sentence-level automatic alignment. However, the system supports a greater number of input formats than the former aligner (see GITS, 2008). Based on a quick test conducted on a text pair in the CODE-UniBO corpus, the application only creates one-to-one bitext link correspondences, which is not the ideal solution in cases where input bitext halves are not always trustworthy mutual translations. Although segmentation options cannot be customized, the alignment quality can be manually improved through an interactive editor.

2.2.3 Commercial GUI alignment tools

The following systems are either commercial standalone alignment applications, i.e. **AlignFactory**,[18] or commercial translation environments including an alignment tool as one of their translation memory creation features, i.e. **SDL Trados Studio 2014** (henceforth, SDL Trados), and **memoQ 2014** (henceforth, memoQ). Unlike the aligners examined so far, these tools align multiple bitexts as sections of a single repository file, which can ultimately be exported to TMX format. Moreover, SDL Trados and memoQ automatically couple input file pairs, which considerably accelerates the file import process. AlignFactory provides the same function.

[18] In the present study we refer to the **AlignFactory Light** version. Information on the various AlignFactory versions is available at: http://www.terminotix.com/ (last visited February 22, 2015).

However, it is not available in the (cheaper) Light version, where files have to be manually paired.

None of these alignment tools needs input file pre-processing or additional resources, even though memoQ users can add term bases to use entries as anchor words and improve the alignment output quality. Beside anchor terms, memoQ's LiveAlign technology relies on unspecified statistical and linguistic algorithms that make use of additional information such as structural and formatting information as well as inline tags (Kilgray, 2015). Unfortunately, the alignment approach used by Align Factory and SDL Trados is not currently known. What we do know is that the three aligners make it possible to configure segmentation rules: AlignFactory relies on eighteen configurable segment filtering options; SDL Trados relies on user-defined segmentation options and its alignment results can be fine-tuned by defining a confidence value threshold; and, finally, memoQ makes use of regular expressions to optimize segmentation results.

Based on our experience, these aligners output one-to-one and many-to-one/one-to-many sentence correspondences (see Appendix A), but none of them deals with partial alignment (*n*-to-zero and/or zero-to-*n* sentence correspondences). AlignFactory and SDL Trados, e.g., merge uncoupled segments with preceding or following segments. On the other hand, memoQ enables users to mark manually deletions and omissions. The alignment can then be rerun leaving marked segments unaltered. All the aligners at issue enable interactive alignment editing. Unlike the other pieces of software, memoQ also attributes a visual confidence value to bitext links by virtue of which users can quickly review and edit the alignment results. As for cross-alignment, none of them is capable of recognizing translation choices that led translators to change the original document structure. Indeed, it must be observed that there is hardly any aligner (if none at all) that supports such feature, at least not among the aligners that were considered in the list in Appendix A. Some of them, however, offer some kind of post-processing function that helps solve partially or totally the cross-alignment problem. For instance, AlignFactory (similarly to Align Assist in Section 2.2.2) allows users to swap cross-aligned segments, and memoQ enables manual creation of cross-links to be left aside during automatic re-alignment (as mentioned above, this feature can be similarly used in the case of *n*-to-zero/zero-to-*n* translation equivalence).

2.3 Evaluation of the aligners' suitability for the present study

In this section, the suitability of the aligners described so far for the present project will be discussed. The aligners will be briefly compared and evaluated with the aim of choosing the one(s) that would possibly meet best the needs of the project, i.e. high-level process automation and accuracy (see also Section 1.4 and the introductory section of Chapter 2).

The limited CLI expertise and the limited time to learn to use command line tools represented the main obstacles that prevented us from selecting one of the aligning algorithms and CLI tools described in Section 2.1. Expertise/time factors also contributed to the specific exclusion of CTK (Ma, 2006), TCA (Hofland & Johansson, 1998), and GMA (Melamed, 1996). To the best of our knowledge, none of these three aligners has been tested on the Italian-English language pair (see Section 2.1). CTK has been originally designed to perform on remote languages, but additional language pairs can be added. Had the author been able to extend CTK's technology to closer languages (i.e. Italian and English), the aligner might still have produced different results from those reported in the literature. On the other hand, TCA and GMA have been tested on several Romance languages (i.e. Portuguese/Spanish and French/Spanish respectively). Due to their close relatedness with the Italian language, it might be hypothesized that these algorithms can be easily extended to our project language pair. However, this operation would have required more time than was available in a study of this type.

Another reason for the exclusion of the aligners described in Section 2.1 is the fact that most of them were designed for the alignment of "clean" parallel texts. Consequently, they might perform poorly on noisy corpora, such as the CODE-UniBO corpus in our study. With the exception of hunalign, the majority of the aligning algorithms and CLI tools have problems recognizing *n*-to-zero/zero-to-*n* correspondences. This is particularly true of early algorithms: in the case of Gale & Church's (1991) and Moore's (2002) algorithms, e.g., the increase in the number of misaligned content in noisy data might be due to the bijectivity and monotonicity constraint at the core of their approach (see Section 2.1). On the other hand, in the case of TCA (Hofland & Johansson, 1998), the fact that one-to-one alignment correspondences are not included in the file to be manually reviewed (see Section 2.1) might have non-negligible consequences on the alignment accuracy. Depending

on the number of anchor words detected in the file pairs to be aligned, the system might find one-to-one correspondences even though particular segments are not mutual translations.

A further limitation of these aligners is the fact that they are not designed as interactive tools. In fact, the absence of a built-in editor makes misalignment correction unlikely to be performed and/or unprofitable, due to the high degree of human effort required. The lack of interactivity also has the consequence that the risk of misalignment cannot be reduced through segmentation rules configuration (an exception is bligner: see Section 2.1). In these cases, input file pre-segmentation at sentence or paragraph level could help increase the output quality and minimize human effort in the post-processing phase. However, the balance between process automation and manual intervention might be affected in the case of poor layout quality.

Similarly to the aligners commented on so far, the quality of the alignment of the commercial tools described in Section 2.2.3 might be very poor in the case of noisy content. As a matter of fact, all of them are unable to deal with *n*-to-zero and zero-to-*n* sentence correspondences. Their multiple post-processing and output configuration features are a practical way to avoid such concrete obstacle. However, all said features involve manual operations that leave little or no space to automation. None of the commercial aligners was therefore chosen for the present project. Nevertheless, it might be worth taking the translation environment memoQ into consideration for further testing in order to establish whether and to what extent a set of institutional academic term base entries added to the software might help improve the initial alignment accuracy.

In the light of the analysis presented in Section 2.2.2, it might be claimed that only two aligners satisfy our selection criteria, i.e. eAlign and InterText. In fact, we expect them to give the best alignment results in terms of both process automation and accuracy. Specifically, they both bundle the aligning algorithm that presumably produces the best results among those described so far (i.e. hunalign) into a user-friendly graphical user interface (see Section 2.2.2). The algorithm enables both aligners to handle one-to-many/many-to-one, one-to-one and *n*-to-zero/zero-to-*n* sentence correspondences (see Section 2.2.1). Hence, it might be hypothesized that these tools perform better than the other aligners on noisy corpora. Moreover, unlike the majority of the aligners described so far, they can handle multiple input files.

This might be extremely useful in the case of a large amount of alignable text pairs (e.g. in the present project more than 3,000 text pairs have to be automatically aligned).

Both tools enable segmentation rules configuration, although non-expert users might find the task easier to perform on eAlign. As a matter of fact, the complexity of regular expressions in InterText's segmentation options might dissuade non-expert users from modifying them (see Section 2.2.2). This might result in decreased alignment success rates, which translates into a higher degree of human effort required in the post-processing phase. In this respect, InterText also requires input files to be segmented at sentence and/or paragraph level in order to yield higher accuracy. This clearly implies a certain level of human intervention, which might increase as the bitext layout quality decreases.

Last but not least, both aligners display a user-friendly interactive editor that enables easy and quick review and correction of the output. They also enable users to create translation memories in TMX format. However, eAlign minimizes post-processing time and human effort by enabling output alignments to be merged into a single TMX file. The latter aspect, together with the more user-friendly options for configuring segmentation rules (see Section 3.2.2), led us to believe that eAlign would strike the best balance between process automation and alignment accuracy. Hence, it was ultimately selected for the alignment of the Italian-ELF web-based academic course descriptions in the CODE-UniBO corpus. The methods used to carry out this task will be discussed in Chapter 3.

Chapter 3 | Methods

Over the past quarter of century, research on sentence alignment has given rise to progressively sophisticated algorithms, resulting in increasingly effective methods to carry out the task (see Chapter 2, and more specifically, Gale & Church, 1991; Brown et al., 1991; Melamed, 1996.; Moore, 2002; Varga et al., 2005. See also Simard et al., 1992; Wu, 1994; Simard & Plamondon, 1998; Tiedemann, 2006). Most of these studies deal with parallel texts with a limited number of omissions and/or insertions, neglecting almost entirely the issue of sentence alignment in noisy data. A couple of exceptions are Chuang & Chang's (2002) and Ma's (2006) works, who evaluate the performance of two sentence aligners on remote languages such as English and Chinese: the first one describes an "especially effective [system] in the case of noisy translations" (Chuang & Chang, 2002: 91), whereas the second one addresses the "robust alignment of potential noisy parallel text" (Ma, 2006: 489). Even in these cases, data noisiness is mainly related to the number of *n*-to-zero and zero-to-*n* bitext link correspondences – or omissions and insertions – in their corpora.

In the light of these considerations, two reasons prompted us to undertake the present study. First of all, to the best of our knowledge, no study on the alignment of noisy corpora at sentence level has been conducted on European languages, and more specifically, on the Italian-English language pair. Secondly, although the CODE-UniBO corpus presents manifold omissions and insertions, its noisiness is also due to several other non-equivalence patterns and structural irregularities, which presumably affect and hinder the corpus automatic alignment process (see Section 1.4.2 for further details). Building on such premises, the present study adopts a four-pronged integrated heuristic approach to the creation of a reference resource for non-native authors/translators through the automatic alignment of a large set of Italian-ELF academic course descriptions in the CODE-UniBO corpus. The four aspects of the approach include the categorization of the CODE-UniBO corpus according to document length similarity (characters) and the selection of a sample to be evaluated (Section 3.1), the estimation of the extent of noisiness/parallelism of the bitexts in the sample (Section 3.2.1), the evaluation of the bitext automatic segmentation and alignment in the sample (Section 3.2.2), and the qualitative analysis of the resulting reference resource in terms of content leverageability (Section 3.2.3). Specifically,

the study is based on three levels of analysis – namely characters, sentences and translation units – and it mainly addresses the following questions:

1. To what extent can one determine the degree/probability of parallelism within a set of bilingual text pairs given as equivalent but not necessarily parallel (i.e. translated)?
2. How does the extent of parallelism of said text pairs relate to their length variation rates?
3. To what extent can a large number of noisy bitexts be automatically aligned at sentence level so as to yield an acceptable error rate with the minimum amount of human effort?
4. Is there any connection between the alignment success rate and the extent of parallelism of the aligned bitexts? And how does this relate to their length variation rates?
5. If we were to create a parallel corpus and/or a translation memory as a result of the automatic alignment of said bitexts, how much of their content could be leveraged?

3.1 CODE-UniBO corpus categorization and sample selection

The answers to these questions required a prior corpus categorization and the subsequent selection of a text pair sample, hence the application of the first prong of our approach. Building on the common assumption that the set of morphological, lexical, and syntactic constructions in an English text usually result in a shorter instance compared to its Italian equivalent, our corpus categorization was based on the variation in characters between the texts in each of the 3,263 extracted pairs (see also Chapter 2). Accordingly, in this preliminary stage of our study the length variation rate of each text pair was computed by dividing the difference between the number of characters of the English text as minuend (NUM_CHAR1) and those of the Italian text as subtrahend (NUM_CHAR2) by the total number of characters of the text pair:

$$\text{LENGTH VARIATION RATE } (\delta) = \frac{\text{NUM\_CHAR1} - \text{NUM\_CHAR2}}{\text{NUM\_CHAR1} + \text{NUM\_CHAR2}}$$

Results showed a span ranging from -92.08% to +52.87%. Specifically, negative values correspond to shorter (presumed) target texts, i.e. English texts with fewer characters than the Italian (presumed) source texts. On the contrary, positive values imply shorter (presumed) source texts, i.e. Italian texts with fewer characters than the English (presumed) target texts. The two extreme instances of said span are provided in Table 1, which shows, from left to right, the number of characters of the English text, the number of characters of the Italian text, the difference in terms of number of characters between the two texts, and their length variation rate.

NUM_CHAR1 [EN]	NUM_CHAR2 [IT]	DIFFERENCE [CHARACTERS]	LENGTH VARIATION RATE
396	9611	-9215	**-92.08554012**
1450	447	1003	**+52.8729573**

Table 1 Variation in characters between text pairs: extreme values.

In order to determine the extent of noisiness/parallelism of the CODE-UniBO corpus, a few preliminary observations had to be made. According to our initial hypothesis, a text pair's degree of parallelism is directly proportional to its length variation rate, i.e. the lower the variation in characters between two texts, the more parallel the text pair. A comprehensive look at the recorded values revealed that the majority of text pairs lay within a variation range of approximately ±35%, which, therefore, corresponded to our sample selection span. In order to test our hypothesis, the latter span was further divided into seven uniform sub-spans, each covering a 10% range. This was done to provide an objective basis for the data collection and to evaluate the corpus at different granularity levels. The purpose is indeed to establish any possible connection between these levels and the results of the analyses conducted on the text pairs (see Section 3.2). The ultimate aim is to identify the length variation ranges that yield the best results in terms of alignment success rate and amount of leverageable content. The text pairs within these categories will be eventually chosen for the automatic creation of a parallel corpus/translation memory (see Section 4.3).

A couple of caveats should be added at this point. First of all, a self-evident phenomenon arises from the division of the sample selection span: since each category shares at least one threshold value with the adjacent one(s) (in italics in Table 2 below), a decision had to be made as to which of the two figures to exclude from them. In line with the above-mentioned hypothesis, we expected text pairs in

the range of ±5% to display the highest degree of parallelism. Moreover, since Italian texts are generally acknowledged to be longer than their English equivalents, we expected text pairs displaying negative values within this range to be more parallel than text pairs with positive length variation rates. For the same reason, we also expected to record a high probability of parallelism for text pairs displaying negative values in the range of (at least) -5%/-15%. The latter range contains indeed the largest number of text pairs in the CODE-UniBO corpus (see Table 4 below).
In the light of these considerations, we identified the ±5% range as the core interval; as such, text pairs displaying the threshold values were included in this sub-span. The threshold values of the remaining sub-spans were established accordingly (Table 2 shows in bold the values included in each range).

	LENGTH VARIATION RANGE
1.	**-35%** $\leq \delta >$ *-25%*
2.	***-25%*** $\leq \delta >$ *-15%*
3.	***-15%*** $\leq \delta >$ *-5%*
4.	***-5%*** $\leq \delta \geq$ *+**5%***
5.	*+5%* $< \delta \geq$ *+**15%***
6.	*+15%* $< \delta \geq$ *+**25%***
7.	*+25%* $< \delta \geq$ +**35%**

Table 2. Length variation ranges.

The other clarification that has to be provided at this point is strictly related to the previous one. Note that the variation in characters between each text pair is expressed as a percentage in decimal format (see, for instance, the values shown in the last column of Table 1). Thus, a further decision had to be made as to whether to round up the obtained decimal values to the next highest or lowest whole number, and use these rounded values to assign text pairs to one range or another. To ensure maximum granularity, we decided to consider rates in decimal format: regardless of the recorded decimal digits, any text pair whose length variation rate fell within the established thresholds was included in the respective sub-span. Some examples are provided in Table 3. For each negative range the lowest rate is shown, whereas the highest rate is provided along with each positive range. An exception is the core interval (±5%), whose threshold range slightly differs from the others insofar as it encompasses both negative and positive values. In this case, both the lowest and highest rates are listed.

LENGTH VARIATION	
RANGE	**RATE**
-35% ≤ δ > -25%	-35.80246914
-25% ≤ δ > -15%	-25.91240876
-15% ≤ δ > -5%	-15.94915949
-5% ≤ δ ≥ +5%	-5.987708516
	5.965362412
+5% < δ ≥ +15%	15.90690517
+15% < δ ≥ +25%	25.0863061
+25% < δ ≥ +35%	35.0617284

Table 3. Lowest and/or highest length variation rates included in each range.

A total number of 3,156 out of 3,263 text pairs were found to belong to the ±35% variation span. On the other hand, a closer look at the number of text pairs belonging to each sub-span supports the assumption mentioned in the introductory paragraph of this section that English texts commonly tend to be shorter than their Italian counterparts. Evidence is provided in Table 4, which shows the number of text pairs for each range and for positive and negative length variation rates. As can be noticed, the majority of bitexts (84.95%) displays negative values, which proves that our assumption is true (at least) for the text type and specific text pairs under consideration (see Section 1.4.2 for further details on the nature of the text pairs in the CODE-UniBO corpus). Nonetheless, careful consideration should be given to the noisy nature of the texts inasmuch as missing sections and/or summarized content in both source and target texts might sometimes influence such analysis.

LENGTH VARIATION RANGE	TEXT PAIRS	NEGATIVE VS. POSITIVE RATES
-35% ≤ δ > -25%	138	2681 (84.95%)
-25% ≤ δ > -15%	327	
-15% ≤ δ > -5%	1111	
-5% ≤ δ ≥ +5%	1412 (1105 + 307)	
+5% < δ ≥ +15%	119	475 (15.05%)
+15% < δ ≥ +25%	39	
+25% < δ ≥ +35%	10	
TOT.	**3,156**	

Table 4. Number of text pairs for each category and for positive and negative length variation rates.

Following the definition and application of the above-mentioned instructions and the categorization of the text pairs according to length variation classes, a

sample was randomly extracted from the corpus. For each listed category, it was initially planned that 15 text pairs would be chosen. However, shortage of text pairs allowed us to include only 10 instances in the range of +25%/+35% (see Table 4). The sample eventually included 100 text pairs, corresponding to roughly 3% of the overall number of instances in our corpus, distributed as follows:

LENGTH VARIATION RANGE	TEXT PAIRS
-35% ≤ δ > -25%	15
-25% ≤ δ > -15%	15
-15% ≤ δ > -5%	15
-5% ≤ δ ≥ +5%	15
+5% < δ ≥ +15%	15
+15% < δ ≥ +25%	15
+25% < δ ≥ +35%	10
TOT.	**100**

Table 5. Sample distribution: bitexts selected for each range.

3.2 Analysis of the sample

As already mentioned, the corpus categorization into length variation ranges and the sample selection represented the first, necessary step of the four-pronged approach adopted in the present study. The core issues of the analysis were addressed instead by the three remaining prongs of the approach, which are aimed at evaluating the degree of parallelism of the text pairs under consideration (henceforth, bitext parallelism) (Section 3.2.1), the success rate of the alignment of said text pairs (Section 3.2.2), and the quality of the resulting sample resource for non-native authors/translators (Section 3.2.3). The possible connection between the data obtained and the document length similarity (i.e. the length variation rates) was also examined, hence the importance of the preliminary approach described in Section 3.1.

3.2.1 Bitext parallelism in the sample

In order to answer the first question addressed by the present study (i.e. to determine the extent of noisiness within a set of bilingual text pairs given as equivalent but not necessarily parallel), the author performed a manual comparison and count of the sentences of each bitext in the sample. For practical reasons, the sample was automatically aligned using the software eAlign (see Sections 2.2.2 and 2.3) and the

alignments were manually post-edited. This enabled us to recognize and count the *sentence pairs*[19] in the various bitexts more easily, quickly and accurately than it would have been possible by opening and comparing the files with a text editor. In other words, at this stage the aligner was only used to support the comparison process. However, a welcome by-product of this step was a manually aligned reference data set against which to compare the results of the software's automatic segmentation in the following stage of our study (see Section 3.2.2 below).

Shifting the focus to the core of the analysis, bitext parallelism was expressed as a percentage: the total number of *parallel* sentences in the source and target texts (TOT_SENT_PARALLEL), divided by the total number of sentences in the text pair (TOT_SENT) – that is to say, the sum of the number of sentences in the English and Italian texts (NUM_SENT1 and NUM_SENT2 respectively):

$$\text{BITEXT PARALLELISM} = \frac{\text{TOT\_SENT\_PARALLEL}}{\text{TOT\_SENT}}$$

Examples taken from the -35%/-25% range are presented in Table 6:

NUM_SENT1 [EN]	NUM_SENT2 [IT]	TOT_SENT	TOT_SENT_ PARALLEL	BITEXT PARALLELISM
18	26	44	31	**0.704545455**
35	39	74	64	**0.864864865**
16	19	35	32	**0.914285714**

Table 6. Analysis of the degree of parallelism of three text pairs in the -35%/-25% range.

Before moving on to describe the two remaining prongs of the approach, a couple of clarifications might prove useful at this point. First of all, a definition of the notion of sentence should be given. For practical reasons, the following constructs were regarded as sentences in the present study:[20]

- any sequence of characters followed by a full stop, except for abbreviations, decimals, alphanumeric formats in lists, domain names (e.g. *http://www.guideweb.unibo.it*), and bibliographic references;
- any sequence of characters followed by a line feed, including metadata (e.g. *###Programma*; cf. Table 3 in Section 1.4.2), text headings and sub-headings;

[19] The term *sentence pairs* may refer to any kind of sentence correspondence (see Section 1.4.3).
[20] See Section 1.4.2 for examples of each construct.

- any sequence of characters followed by a colon, if the subsequent non-whitespace character was either a capital letter, a bulleted/numbered/simple list element, a domain name, or a line feed;
- any sequence of characters followed by a semicolon, if the subsequent non-whitespace character was either a capital letter, a bulleted/numbered/simple list element, or a line feed;
- in the case of missing punctuation, any sequence of characters preceded by a capital letter and/or any sequence of characters underlying a consistent and coherent message;
- list items and isolated domain names, that is to say, URLs preceded by a colon and/or a line feed;
- items in bibliographic reference lists, irrespective of their punctuation.

Secondly, the concept of parallel sentences as defined in our analysis should be clarified. Due to the nature of the texts in the CODE-UniBO corpus, which are hardly ever close equivalents of each other (see Section 1.4.2), a recall-maximizing strategy was adopted: any significant amount of *translated* content that could be leveraged (under some circumstances) was considered as parallel. In other words, if two or more sentences in the source and target texts shared substantial reusable information from a resource-oriented perspective (i.e. machine translation systems, computer-aided tools), they were included in the parallel sentence count. For instance, example 1 in Table 7 was considered to be parallel, while example 2 was not.

	SOURCE CONTENT [IT]	**TARGET CONTENT [ELF]**
1.	Lo studente, alla fine del corso, deve dimostrare la capacità di comprendere concetti e argomenti relativi alle scienze sociali, alle scienze politiche, in generale, e alla politica internazionale e le sue notizie espresse in tedesco e dovrà inoltre essere in grado di esprimersi sugli stessi argomenti in lingua tedesca ad un livello non inferiore a A2 e, idealmente, a B1 (secondo il Quadro di Riferimento del Consiglio d'Europa). Per quanto concerne la lingua	At the end of the course students should reach a minimum level of A2 (Council of Europe framework) for spoken German, while the ideal level is B1, and a minimum level of B1 for reading comprehension – ideal level B2.

	scritta lo stesso studente deve essere capace di leggere e comprendere testi complessi a un livello non inferiore a B1 e, idealmente, al livello B2.	
2.	La verifica dell'apprendimento avviene attraverso un esame che accerti le conoscenze richieste mediante una prova scritta della durata di 90 minuti ed una prova orale sul programma svolto.	Written text and oral examination.

Table 7. Example of parallel and non-parallel sentences.

Although both target sentences lack some content information and the two sentences in the second example share a minimum degree of translation equivalence, the target sentence in example 2 was considered to be a summary rather than a translation of the source sentence. These cases were ignored because they are not accurate enough for our ultimate aim, i.e. to provide a reference resource for assisting non-native authors/translators in the production of web-based academic course descriptions in ELF (see Section 1.4 for further details). It should be noted, however, that although every effort was made to ensure objectivity, this analysis bears a minimum level of subjectivity. Indeed, to some extent the evaluation of the extent of translation equivalence between two or more sentences always implies subjective estimates, especially when one deals with noisy data.

Besides exemplifying our concept of parallel and non-parallel sentences, these instances might also be useful to explain the manual sentence count performed on the text pair sample in this particular stage of our study. As can be noticed, in the first example translation equivalence takes the form of a many-to-one parallel correspondence, where a single English sentence corresponds to two Italian sentences. In this case, three sentences were counted as parallel. The same applied to any type of parallel correspondence, be it 1:1 (two sentences), 1:2 (three sentences), 2:2, 1:3, 3:1 (four sentences), and so forth. As for example 2 in Table 7, since it was deemed that there was no translation equivalence between the two sentences, they were not included in the count.

3.2.2 Bitext automatic segmentation and alignment in the sample

Not only did the sentence count serve as the basis for the analysis of the degree/probability of parallelism in the sample, it was also used for the evaluation of its segmentation and alignment. In this stage of the study, our purpose was to find and evaluate a machine-driven method that would enable us to align profitably a large number of noisy bitexts at sentence level with the minimum amount of human effort (see question 3 at the beginning of this chapter). The ultimate aim was to investigate any possible connection among the results of this analysis and those described so far that would enable us to trace the bitext alignment accuracy to the variation in characters between the two bitext halves (see question 4). While the aligner used for the present project (i.e. eAlign) was chosen because it provided a satisfactory balance between process automation and output accuracy (see Section 2.3), it is not necessarily the case that it was able to align correctly every instance that a human scorer evaluates as parallel. Hence the need to keep the two levels (parallel vs. alignable/aligned) separate.

In this phase, the text pairs were automatically realigned using eAlign. This time, however, the output was not post-edited. While no correction was performed on the alignment, segmentation rules were fine-tuned with the aim of obtaining the closest results to the bitext segmentation in the manually created reference alignments mentioned in Section 3.2.1. This is because a direct connection between segmentation and alignment results was hypothesized whereby a perfectly segmented text pair is more likely to return the lowest alignment error rate. It should be stressed that the fine-tuning of the segmentation rules also requires a certain amount of human effort. However, it is a one-stage operation that strikes a good balance between no intervention at all and manual alignment/misalignment correction. In practical terms, provided that an expert user has a general knowledge of the structure of the bitexts, he/she would take a maximum of 20/30 minutes to carry out the task. On the other hand, a non-expert user would take between 30 and 90 minutes to learn to configure the software's segmentation rules and to apply them profitably.[21] Conversely, based on our experience, we estimate that (non-)expert users would take an average time of 15/20 minutes to correct manually the

[21] These estimates refer to the use of the software eAlign (see Section 2.2.2).

segmentation and alignment of a single item.[22] In the case of a large set of text pairs, the latter operation is clearly unprofitable both in terms of time and human effort. For obvious reasons, in this case the fine-tuning of the segmentation rules is therefore the best option.

A full account of the specific segmentation rules used in our project is provided in Tables 8, 9 and 10. Table 8 shows the strings that eAlign considered as the end of the sentence in both Italian and English texts (the only string added to the original set is the semicolon).

SENTENCE ENDING STRINGS				
...")	?")	.”)	?"	.’
. . .	!")	...	.”	?’
..."	…")	!”)	.)	!’
...”	…?”	.’”)	!)	…’
...”	…!”	!’”)	.]	.
...’	.’”)	?’”)	!]	!
...)	.).	…”	?]	?
...?	.):	…)	!”	…
...!	...	."	?”	:
.")	?”)	!"	?)	;

Table 8. Sentence ending strings in Italian and English texts.

On the other hand, Tables 9 and 10 show the strings that eAlign considered as exceptions, i.e. those strings that the software would skip even if they contained one of the sentence ending strings listed in Table 8. In this case, two different sets of exceptions were configured for Italian and English texts (Tables 9 and 10 respectively) according to their different linguistic and structural nature. It is crucial to note that these segmentation rules are not recommended for every text type and language pair, since they were moulded to the specific structure of the texts in the CODE-UniBO corpus. For both the Italian and English language, we present the built-in options and the exceptions added by the author in the same order as in the settings file (top-down order in Tables 9 and 10). The author modified several built-in strings in the Italian settings file through the insertion of the wildcard “*” (in bold in Table 9). The same changes were reflected in the strings added by the author in the English settings file (in bold in Table 10), since the latter presented fewer built-in exceptions than the Italian counterpart. In this respect, readers might notice some

[22] The estimation of the time/effort required to correct the automatic segmentation and alignment of a text pair depends on the length of the texts and the quality of the raw output.

wildcards and regular expressions among the listed strings; therefore, a brief explanation of their use follows:[23]

- the wildcard "*" as the first character of the string means that the exception is case sensitive (e.g. "*Mme." recognizes as an exception "**Mme.** Bovary", but not "progra**mme.**");
- the characters "^#" stand for any digit (e.g. "^#.^#." handles numbers in the form of "2.2." as an exception);
- "[a-z]" means any lower-case letter;
- "[A-Z]" stands for any capital letter;
- the backslash character "\" is used to treat any subsequent character as a literal in regular expressions (e.g. the list item "*1. Introduction to Clinical psychology*" is not segmented after the full stop since the software recognizes the regex "*^#\. [A-Z]" as an exception).

EXCEPTIONS [IT]					
BUILT-IN OPTIONS					
*[A-Z]\.[A-Z]	i.e.	No. ^#	Adj.	Hosp.	***Res.**
U.S.	i. e.	U.S.	Adm.	Insp.	Rev.
Gen. Non-	.org	*et al\. [a-z]	Adv.	Lt.	***Rt.**
jan.	.net	R.O.C.	Asst.	MM.	Sen.
feb.	.com	*No\. [0-9]	Bart.	MR.	Sens.
mar.	www.	*Inc\.	Bldg.	MRS.	Sfc.
*PC:	L. Ron	Ltd.	Brig.	***MS.**	Sgt.
*LRH:	^#.^t	Inc.	Bros.	Maj.	Sr.
apr.	*\.\.\. [a-z]	*\.\.\. [a-z]	Capt.	Messrs.	***St.**
jun.	*St\.	p.m.	Cmdr.	Mlle.	Supt.
jul.	Mr.	a.m.	Col.	***Mme.**	Surg.
aug.	Mrs.	.^#	Comdr.	Mr.	vs.
sep.	*etc\.\) [a-z]	^?—	Con.	Mrs.	i.e.
sept.	*etc\. [a-z]	^?"—	Corp.	***Ms.**	rev.
oct.	**^#.^#.**	*\: [a-z]	Cpl.	Msgr.	e.g.
nov.	*\?" [a-z]	^?)—	DR.	Op.	No ^#.
dec.	*\!" [a-z]	^? —	Dr.	***Ord.**	Nos ^#.
^#. R	Dr.	*…" [a-z]	Drs.	Pfc.	Art ^#.
a.d.	etc.,	*\… [a-z]	***Ens.**	***Ph.**	Nr.
b.c.	etc.),	*\.\.\. [a-z]	Gen.	Prof.	pp ^#.
a. d.	Ph.D.	*\. \. \. [a-z]	Gov.	Pvt.	
b. c.	e.g.	dott.	Hon.	Rep.	
*\! [a-z]	e. g.	Sig.	Hr.	Reps.	
STRINGS ADDED					

[23] Further details are available in the software's Help section.

(C.I.)	*eds\.	* N.	*^# \. [A-Z]	*^# \)[a-z]	\: \"
eg.	*Eds\.	num.	*^# \. [a-z]	*[A-Z]\?[A-Z]	\; \+
prof.	a.a.	Num.	*^# \.[A-Z]	*[A-Z]\?[a-z]	\: \(
*p.	A.A.	es.	*^# \.[a-z]	*[a-z]\?[a-z]	\.\)
pp.	vol.	chap.	*^#\) [A-Z]	*[a-z]\?[A-Z]	*\; [a-z]
pag.	voll.	a c. di	*^#\) [a-z]	*[a-z]\:\/	* [A-Z]\.
pagg.	*Art.	&	*^#\)[A-z]	.it	
cap.	N.B.	*^#\. [A-Z]	*^#\)[a-z]	*\.[A-Z]\.	
capp.	tot.	*^#\. [a-z]	*^# \) [A-Z]	*[a-z]\.[a-z]	
*ed\.	Tot.	*^#\.[A-z]	*^# \) [a-z]	*[a-z]\:[a-z]	
*Ed\.	* n.	*^#\.[a-z]	*^# \)[A-Z]	*\: [A-Z]	

Table 9. Strings considered as exceptions in the segmentation of the Italian texts.

EXCEPTIONS [EN]					
BUILT-IN OPTIONS					
*[A-Z]\.[A-Z]	sep.	i. e.	*etc\. [a-z]	*et al\. [a-z]	*\: [a-z]
U.S.	sept.	.org	**^#.^#.**	R.O.C.	^?)–
Gen. Non-	oct.	.net	*\?" [a-z]	*No\. [0-9]	^? –
jan.	nov.	.com	*\!" [a-z]	*Inc\.	*…" [a-z]
feb.	dec.	www.	Dr.	Ltd.	*\… [a-z]
mar.	^#. R	L. Ron	etc.,	Inc.	*\.\.\. [a-z]
*PC:	a.d.	^#.^t	etc.),	*\.\.\. [a-z]	*\. \. \. [a-z]
*LRH:	b.c.	*\.\.\. [a-z]	Ph.D.	p.m.	
apr.	a. d.	*St\.	e.g.	a.m.	
jun.	b. c.	Mr.	e. g.	.^#	
jul.	*\! [a-z]	Mrs.	No. ^#	^?—	
aug.	i.e.	*etc\.\) [a-z]	U.S.	^?"—	
STRINGS ADDED					
dott.	Gen.	Pfc.	No ^#.	N.B.:	*^# \) [a-z]
Sig.	Gov.	***Ph.**	Nos ^#.	tot.	*^# \)[A-Z]
Adj.	Hon.	Prof.	Art ^#.	Tot.	*^# \)[a-z]
Adm.	Hr.	Pvt.	Nr.	num.	*[A-Z]\?[A-Z]
Adv.	Hosp.	Rep.	pp ^#.	Num.	*[A-Z]\?[a-z]
Asst.	Insp.	Reps.	(C.I.)	chap.	*[a-z]\?[a-z]
Bart.	Lt.	***Res.**	eg.	a c. di	*[a-z]\?[A-Z]
Bldg.	MM.	Rev.	prof.	&	*[a-z]\:\/
Brig.	MR.	***Rt.**	pp.	*^#\. [A-Z]	.it
Bros.	MRS.	Sen.	pag.	*^#\. [a-z]	* [A-Z]\.[A-Z]\.
Capt.	***MS.**	Sens.	pagg.	*^#\.[A-z]	*[a-z]\.[a-z]
Cmdr.	Maj.	Sfc.	cap.	*^#\.[a-z]	*[a-z]\:[a-z]
Col.	Messrs.	Sgt.	capp.	*^# \. [A-Z]	*\: [A-Z]
Comdr.	Mlle.	Sr.	*Ed\.	*^# \. [a-z]	\: \"
Con.	***Mme.**	***St.**	*eds\.	*^# \.[A-Z]	\; \+
Corp.	Mr.	Supt.	*Eds\.	*^# \.[a-z]	\: \(
Cpl.	Mrs.	Surg.	a.a.	*^#\) [A-Z]	\.\)
DR.	***Ms.**	vs.	A.A.	*^#\) [a-z]	*\; [a-z]
Dr.	Msgr.	i.e.	vol.	*^#\)[A-z]	* [A-Z]\.
Drs.	Op.	rev.	voll.	*^#\)[a-z]	

***Ens.**	***Ord.**	e.g.	*Art.	*^# \) [A-Z]	

Table 10. Strings considered as exceptions in the segmentation of the English texts.

Adopting a similar method to the one in Section 3.2.1, a count was performed of the number of sentences segmented in the same way as the human segmentation. For clarification purposes, Table 11 provides an example of incorrect segmentation due to missing punctuation. Specifically, the left column shows the sentences as they were segmented by the author, whereas in the right column is their machine-segmented counterpart.

MANUAL SEGMENTATION	AUTOMATIC SEGMENTATION
Spiral, code-based and/or model-driven software development processes	Spiral, code-based and/or model-driven software development processes Object based software systems Introduction to the working tools: the Eclipse framework Introduction to UML Introduction to the Design Patterns Techniques and methlogies for continous integration and cooperating working Usage of software components (OSGi and Eclipse plugin)
Object based software systems	
Introduction to the working tools: the Eclipse framework	
Introduction to UML	
Introduction to the Design Patterns	
Techniques and methlogies for continous integration and cooperating working	
Usage of software components (OSGi and Eclipse plugin)	

Table 11. Example of incorrect segmentation due to missing punctuation.

In this case, all the seven sentences were excluded from the count, for obvious reasons. A similar procedure was applied to sentences in rows 2 and 4 in the left column of Table 12, which integrates the previous one by comparing the manual and automatic segmentation and alignment of the English sentences listed in Table 11 and their Italian equivalents. The automatically segmented sentences in rows 1.1 and 3.1 (in bold in the left column of Table 12) were the only instances included in the segmentation count.

MANUAL SEGMENTATION AND ALIGNMENT			
1	Processi di sviluppo del software a spirale model-driven e/o code-based .	**1**	Spiral, code-based and/or model-driven software development processes
2	Richiami sulla costruzione di sistemi software ad oggetti,	**2**	Object based software systems
3	Introduzione agli strumenti di lavoro: il framework Eclipse.	**3**	Introduction to the working tools: the Eclipse framework
4	Concetti fondamentali del linguaggio UML.	**4**	Introduction to UML
5	Design pattern ed esempi di applicazione.	**5**	Introduction to the Design Patterns
6	Tecniche e metodologie di	**6**	Techniques and methlogies for

	integrazione continua e collaudo nel lavoro singolo e di gruppo.		continous integration and cooperating working
7	Utilizzo di componenti software (OSGi e Eclipse plugin).	**7**	Usage of software components (OSGi and Eclipse plugin)

AUTOMATIC SEGMENTATION AND ALIGNMENT			
1.1	**Processi di sviluppo del software a spirale model-driven e/o code-based .**	**1.1**	
2.1	Richiami sulla costruzione di sistemi software ad oggetti, Introduzione agli strumenti di lavoro: il framework Eclipse.	**2.1**	
3.1	**Concetti fondamentali del linguaggio UML.**	**3.1**	
4.1	Design pattern ed esempi di applicazione. Tecniche e metodologie di integrazione continua e collaudo nel lavoro singolo e di gruppo. Utilizzo di componenti software (OSGi e Eclipse plugin).	**4.1**	Spiral, code-based and/or model-driven software development processes Object based software systems Introduction to the working tools: the Eclipse framework Introduction to UML Introduction to the Design Patterns Techniques and methlogies for continous integration and cooperating working Usage of software components (OSGi and Eclipse plugin)

Table 12. Comparison of manual and automatic segmentation and alignment.

Not only does this example clarify the method used in the segmentation analysis, it also provides information on our approach to the evaluation of the sample alignment success rate. Such rate was computed based on the count of the number of sentences correctly aligned by the software. Although the method adopted for this analysis is similar to the previous one, it differs from it insofar as the resulting automatic data set was not compared to any of the reference alignments mentioned in Section 3.2.1. To avoid conducting a subjective analysis, the computation of the alignment success rate relied on the results of the software's automatic segmentation. Moreover, since our main purpose was to determine the amount of text content information that could be reused and leveraged, recall was preferred over precision. What this means is that, regardless of the sentence correspondences in the output and independently from any possible extra context, one or more sentences were considered to be correctly aligned provided that a degree of translation equivalence was observed. Likewise, in the case of *n*-to-zero and zero-to-*n* sentence correspondences, the alignment of one or more sentences in

one bitext half was deemed correct if the software recognized no equivalence to any sentence in the other bitext half. For instance, in the previous example (Table 12), six sentences out of fourteen were considered to be correctly aligned, all of which can be found in row 4.1. Due to the reasons outlined above, the target sentences under consideration were all grouped together leading to a similar, albeit not comprehensive, result in the left-hand side of the grid (corresponding to source texts). Although the four source sentences in rows 1.1 to 3.1 and the first four target sentences in row 4.1 were misaligned, the remaining instances could *de facto* be leveraged, hence their inclusion in the number of correctly aligned sentences. Had the first four sentences in the source side of the grid been originally presented as a one-to-zero parallel correspondence, they would also have been considered to be correctly aligned.

Bearing these considerations in mind and moving to the actual computation of the segmentation and alignment success rates, these were expressed as the percentage of the total number of correctly segmented and aligned sentences (TOT_SENT_SEGM and TOT_SENT_ALIGN respectively) divided by the total number of sentences in the bitext (TOT_SENT):

$$\text{SEGMENTATION ACCURACY} = \frac{\text{TOT\_SENT\_SEGM}}{\text{TOT\_SENT}}$$

$$\text{ALIGNMENT ACCURACY} = \frac{\text{TOT\_SENT\_ALIGN}}{\text{TOT\_SENT}}$$

By way of example, below are the recorded data for the three previously listed text pairs (Table 6) in the -35%/-25% range:

TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT_ ALIGN	ALIGNMENT ACURACY
44	27	**0.613636364**	33	**0.75**
74	68	**0.918918919**	69	**0.932432432**
35	35	**1**	35	**1**

Table 13. Analysis of the segmentation and alignment success rate of three text pairs in the -35%/-25% range.

It should be underlined that no pre-processing was performed on the sample, although a few structural adjustments would have presumably produced better results in both analyses. This decision was motivated by a practical consideration:

indeed, a change to the layout of one or more texts would have modified the number of characters of the adjusted text pairs, altering the essence and quality of the present study.

3.2.3 Resource-oriented qualitative analysis

The last prong of the approach adopted in our study implied a shift in the object of analysis. While the comparisons so far have been based on the number of characters and sentences respectively, the analysis of the content information that could be leveraged was calculated on the basis of the number of correct translation units sent to a translation memory created ad hoc. In other words, in the last stage of our study, the quality of a translation memory created through automatic alignment was evaluated (see question 5 at the beginning of this chapter). In broad terms, the amount of *parallel* bitext links within an aligned parallel corpus was also evaluated (i.e. the number of correct bitext links that did not contain *n*-to-zero and zero-to-*n* sentence correspondences). For practical purposes, we will only refer to *translation memory* (TM) quality/accuracy and correct *translation units* (TUs).

As in the case of the analysis of the bitext parallelism in the sample (Section 3.2.1), the correctness of the various translations was not evaluated in linguistic terms. It should be noted, indeed, that the concept of quality assurance as defined in the present study does not imply the quality of the language used by the authors/translators, but it rather refers to the quantity of reusable translation units within the translation memory. More specifically, since recall was preferred over precision (in line with the previous approach), any translation unit that comprised at least a one-to-one parallel correspondence was deemed correct. Moreover, in this phase, *n*-to-zero and zero-to-*n* bitext link correspondences were not taken into account, since they are automatically discarded by the software in the creation process of the translation memory. In this respect, this analysis differs from the one in Section 3.2.2 (i.e. bitext automatic alignment accuracy), which also entailed the leveraging of the content of the text pairs. By way of example, although the previously listed translation unit (see Table 12, row 4.1) presents four extra sentences, it also comprises three one-to-one parallel correspondences, hence its leverageability and inclusion in the number of correct instances. Conversely, since the automatically aligned sentences in row 2.1 of Table 14 are not mutual

translations, the translation unit does not meet our analysis requirements, and therefore, it is ignored.

MANUAL SEGMENTATION AND ALIGNMENT			
1	L'esame consta di una prova orale preceduta da una prova scritta, da sostenersi entrambe nel medesimo appello.	**1**	
2	Limitatamente agli appelli di esami previsti per l'anno accademico di frequenza, sono esonerati dal sostenere la prova scritta gli allievi che abbiano superato positivamente le prove di accertamento intermedie, programmate durante il ciclo di insegnamento.	**2**	
3		**3**	At the end of the course the students know the basic concepts of the solid mechanics and the methodologies for the structural analysis.

AUTOMATIC SEGMENTATION AND ALIGNMENT			
1.1	L'esame consta di una prova orale preceduta da una prova scritta, da sostenersi entrambe nel medesimo appello.	**1.1**	
2.1	Limitatamente agli appelli di esami previsti per l'anno accademico di frequenza, sono esonerati dal sostenere la prova scritta gli allievi che abbiano superato positivamente le prove di accertamento intermedie, programmate durante il ciclo di insegnamento.	**2.1**	At the end of the course the students know the basic concepts of the solid mechanics and the methodologies for the structural analysis.

Table 14. Example of misaligned translation unit: comparison of manual and automatic segmentation and alignment.

Using a similar formula to those described for the previous computations, the evaluation of the quality of the ad hoc translation memory was calculated by dividing the number of correct translation units (TOT_TU_CORRECT) by the total number of translation units sent to the translation memory (TOT_TU):

$$\text{TM ACCURACY} = \frac{\text{TOT\_TU\_CORRECT}}{\text{TOT\_TU}}$$

Table 15 shows, by way of example, the results of the analysis conducted on the three previously listed text pairs (Tables 6 and 13). As in the previous analyses, the recorded data are expressed as percentages in decimal format.

TOT_TU	TOT_TU_CORRECT	TM ACCURACY
12	9	**0.75**
28	27	**0.964285714**
13	13	**1**

Table 15. Analysis of the correct translation units sent to the translation memory by three text pairs in the -35%/-25% range.

Complete results along with the answers to the questions addressed by the present study will be thoroughly discussed in Chapter 4.

Chapter 4 | Results and discussion

In this chapter, the data obtained from each analysis described in Section 3.2 will be presented and discussed vertically and horizontally, i.e. at *sample* and *range* level. Specifically, Section 4.1 will present the results of the analyses of the probability of parallelism in the sample (Section 4.1.1), the bitext automatic segmentation and alignment (Section 4.1.2), and the leverageability of the resulting resource (Section 4.1.3). On the other hand, Section 4.2 will discuss the results presented in Section 4.1, answering the questions addressed by the present study (cf. Chapter 3). Data obtained from each analysis will eventually help us identify in Section 4.3 the specific bitexts that would be more suitable in terms of document similarity for the creation of an automatically aligned reference resource for non-native authors and translators working with English in the institutional academic domain. Finally, building on the discussion of the results, Section 4.4 will outline several future perspectives related to this study and the literature reviewed in Chapter 1.

4.1 Results

4.1.1 Bitext parallelism in the sample

The first prong of our heuristic approach led to the categorization of a corpus of Italian-ELF academic course descriptions (i.e. the CODE-UniBO corpus) according to document length similarity (characters) and to the subsequent selection of a sample. This served as a useful starting point for the analyses carried out in the present study (cf. Chapter 3, and specifically, Section 3.1).

The second prong of the approach addressed two questions. To begin with, the study investigated the degree/probability of parallelism of a set of bilingual text pairs given as equivalent but not necessarily parallel (i.e. translated). Our analysis found that the overall degree/probability of parallelism of the bitexts in the sample (henceforth, bitext parallelism) is rather high. For space purposes, it might be useful to report exclusively the *average* results for each range in the various sections of the chapter. In this respect, Table 1 shows the average data obtained from the analysis of the bitext parallelism in the sample, distributed according to length variation ranges. Complete data recorded for this analysis are presented in Appendix B, where the length variation rates and the probabilities of parallelism of each text pair are

reported in bold. In particular, the probabilities of parallelism were computed as the percentage of the difference between the *parallel* sentences[24] and the total number of sentences in the bitexts (see Section 3.2.1). Appendix B also reports in bold the average parallelism data at the end of each range.

LENGTH VARIATION RANGE	BITEXT PARALLELISM [AVG]
$-35\% \leq \delta > -25\%$	0.850598909
$-25\% \leq \delta > -15\%$	0.900272415
$-15\% \leq \delta > -5\%$	0.954722531
$-5\% \leq \delta \geq +5\%$	**0.957580269**
$+5\% < \delta \geq +15\%$	0.897957767
$+15\% < \delta \geq +25\%$	0.897941311
$+25\% < \delta \geq +35\%$	0.720750859

Table 1. Bitext parallelism for each range.

With the exception of the texts in the +25%/+35% length variation range, displaying 0.7207 average probability of translation equivalence, the average values obtained through the count of parallel sentences go from 0.8506 (in the -35%/-25% range) to 0.9576 (in the -5%/+5% range). It should be remembered that a recall-maximizing strategy was adopted in the present study: any significant amount of translated content that could be leveraged was considered as parallel (see Section 3.2.1). Thus, the results presented and discussed in the whole chapter evaluate recall only, not precision. It should be also remembered that the ±5% range was regarded as the core interval (in bold in Table 1) and it was hypothesized that this range would display the highest bitext parallelism among the ranges set for this study (see Section 3.1). Its 0.9576 probability of translation equivalence confirms this hypothesis (see Table 1).

The results presented so far imply that the average probability of translation equivalence among bitexts increases as the length variation rate within said bitexts approaches 0 (see Table 1). Results show indeed an upward trend from the -35%/-25% range to the core interval (i.e. ±5%), where the probability of parallelism reaches its peak. Conversely, a downward trend is observed from the latter interval to the +25%/+35% range (see also Table 1). In practical terms, bitext parallelism data take the form of a parabolic trend from the lowest negative range to the highest positive range. This parabolic trend is better exemplified by the chart in Figure 1,

[24] See Section 3.2.1 for a definition of the notion of sentence and parallel sentences.

where the average length variation rate for each range is displayed on the horizontal axis and the average probabilities of parallelism are shown on the vertical axis.

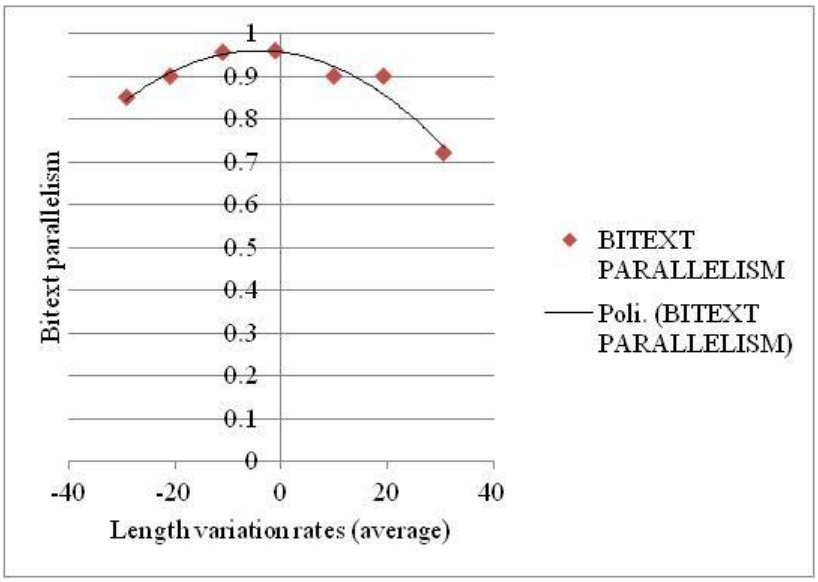

Figure 1. Parabolic trend in the analysis of the bitext parallelism at sample level.

Findings also reveal that it is more likely to find 100% parallelism in bitexts which are highly similar in terms of number of characters. Specifically, the whole sample includes six bitexts (6%) that display full translation equivalence (in bold in Table 2 below). In line with the results presented so far, all of them were found in the -15%/+5% range (see also Appendix B).

LENGTH VARIATION RANGE	100% PARALLEL BITEXTS
$-35\% \leq \delta > -25\%$	0/15 (0%)
$-25\% \leq \delta > -15\%$	0/15 (0%)
$-15\% \leq \delta > -5\%$	**2/15 (13.3%)**
$-5\% \leq \delta \geq +5\%$	**4/15 (26.7%)**
$+5\% < \delta \geq +15\%$	0/15 (0%)
$+15\% < \delta \geq +25\%$	0/15 (0%)
$+25\% < \delta \geq +35\%$	0/10 (0%)
TOT.	**6/100 (6%)**

Table 2. Distribution of 100% parallel bitexts.

On the other hand, it should be also pointed out that as the length variation rate increases (i.e. becomes greater than or close to 0), it is more likely to find English bitexts, i.e. ELF-ELF text pairs. Results reveal indeed that some of the (presumed) Italian-ELF bitexts are actually composed of two similar/identical texts in English. This is particularly the case for text pairs with extremely low or null length variation

rates and text pairs that contain longer English texts than their presumed Italian equivalent. Accordingly, findings show that a total number of 21 bitexts out of 100 (21%) in the sample does not contain an Italian source text and that the majority of English bitexts display rates in the range of -5%/+35%. Table 3 shows in bold the distribution of these items among the ranges (see also Appendix B).

LENGTH VARIATION RANGE	ENGLISH BITEXTS
$-35\% \leq \delta > -25\%$	0/15 (0%)
$-25\% \leq \delta > -15\%$	**1/15 (6.7%)**
$-15\% \leq \delta > -5\%$	0/15 (0%)
$-5\% \leq \delta \geq +5\%$	**3/15 (20%)**
$+5\% < \delta \geq +15\%$	**6/15 (40%)**
$+15\% < \delta \geq +25\%$	**8/15 (53.3%)**
$+25\% < \delta \geq +35\%$	**3/10 (30%)**
TOT.	**21/100 (21%)**

Table 3. Distribution of English bitexts in the sample.

In order to explain the nature of these types of items, Table 4 shows an example of manually aligned bitext with English as a common language of communication on both Italian and English web pages. The most probable reason for the publication of English content on web pages where Italian should be the language of communication is that the course is taught in English.

ITALIAN WEB PAGE [ELF]	ENGLISH WEB PAGE [ELF]
###Corso di laurea: TOURISM ECONOMICS AND MANAGEMENT / ECONOMIA E MANAGEMENT DEL TURISMO ###Titolo: E-COMMERCE AND WEB MANAGEMENT IN TOURISM C.I.	###Corso di laurea: TOURISM ECONOMICS AND MANAGEMENT / ECONOMIA E MANAGEMENT DEL TURISMO ###Titolo: E-COMMERCE AND WEB MANAGEMENT IN TOURISM C.I.
###Programma	###Programma
- Internet distribution of hospitality services	- Internet distribution of hospitality services
- Channel management and profitabilty	- Channel management and profitabilty
- GDS, chain CRS and Meta Search Engines	- GDS, chain CRS and Meta Search Engines
- Flash Sales and Private Sales	- Flash Sales and Private Sales
- Revenue Management practices and models	- Revenue Management practices and models
- Competitive set performance through ADR, OR, RevPAR, RGI and other kpi	- Competitive set performance through ADR, OR, RevPAR, RGI and other kpi
- The impact of Revenue Management on the P&L	- The impact of Revenue Management on the P&L
- Simulating revenue management activities through Opera PMS -	- Simulating revenue management activities through Opera PMS -

(Lab.)	(Lab.)
###Metodi	###Metodi
###Tipo	###Tipo
###Obiettivi	###Obiettivi
	At the end of the course the student will manage a general knowledge of most used web channels for distributing hospitality and other tourism services and will be able to evaluate strategic alternatives on market positioning and distribution, according to profitability.
	Further more, the student will learn how to start up, control and adapt a pricing policy, through the study of most relevant performance kpi and the use of a PMS (Opera).
###Supporti	###Supporti

Table 4. Example of English bitext.

A common feature of these bitexts is that at least one of the bitext halves presents missing sections (e.g. see the Italian side of the grid in Table 4) or cross-references to the Italian or English corresponding web page. Had the two halves been identical, it would have been easier to identify these items from the mere calculation of their length variation rate (i.e. 0). However, since they display a different number of characters, it is difficult to detect them automatically by simply looking at their length variation rates. The extent of parallelism of these types of bitexts is considered to be equal to 0 in our analysis (see Appendix B). Moreover, these items were excluded from the computation of the average results of each analysis at both range and sample level. In other words, the average data outlined so far and those resulting from the other analyses only include results from pairs of Italian and ELF texts. This is because we are only interested in investigating parallelism, aligner's performance and content leverageability across texts in two different languages. It would indeed be useless to align two texts in the same language, for obvious reasons. In this respect, it is as if we identified and excluded English bitexts from the corpus through a priori language detection, which is a relatively simple task that can be performed independently from alignment.

The same approach was applied to bitexts that comprised machine-translated target (i.e. English) texts. In our sample, 1 out of 100 text pairs (1%) presents machine-translated content. In fact, the target text under consideration displays a series of typical errors made by MT systems. Whether or not the raw output of the

MT system was post-edited or published without corrections, the quality of language in the target text is poor. Although the purpose of the present study is not to evaluate the quality of language used in the CODE-UniBO corpus (see Sections 1.4.2 and 3.2.3), it would be useless to align this and other similar texts. Table 5 shows some of the errors which a non-native author/translator might find if the texts were to be aligned and reused for the production of ELF academic course descriptions.

SOURCE SENTENCES [IT]	TARGET SENTENCES [EN]
Prova del primo teorema fondamentale della valutazione.	Try the first fundamental theorem of the evaluation.
Secondo teorema della valutazione.	according to theorem of the evaluation.
Incompletezza del modello trinomiale.	Incompleteness of the model trinomial.
Opzioni Americane e strategie d'esercizio anticipato.	options American and strategies of early exercise.
Viene lasciata la possibilità di scelta di almeno un argomento.	It is left to the choice of at least one argument.

Table 5. Examples of machine-translated sentences.

Since each sentence in the source text has an equivalent in the target text, this text pair is considered to be 100% parallel. Thus, it is comprised among the 6 bitexts having full translation equivalence in Table 2 above. However, this text pair is excluded from the average degree of parallelism of the corresponding range, it is not examined in the other analyses, and therefore, it is not included in the computation of the respective average results. It would indeed be unprofitable for non-native authors/translators to use the aligned bitext as a reference material. Despite the exclusion of both English and machine-translated bitexts from the main computations and analyses, the average of the whole set of bitexts is also reported in Appendix B for the sake of thoroughness (cf. AVG(0s)). However, as already mentioned, the whole chapter (including the discussion presented so far) focuses exclusively on the results of Italian-ELF bitexts, i.e. those listed in Table 1 and along the AVG value in Appendix B.

The other question addressed by this second prong of the approach concerned the relation between bitext parallelism in the sample and document similarity (in terms of variation in characters). Findings in this respect will be discussed in Section 4.2.1.

4.1.2 Bitext automatic segmentation and alignment in the sample

The third prong of our approach explored the extent to which the automatic alignment of the text pairs in the CODE-UniBO corpus at sentence level would yield an acceptable success rate. The aim was therefore to evaluate the performance of the aligner chosen for the present study (i.e. eAlign; see Section 2.3) and to find a profitable automated method to align a large number of noisy bitexts with the minimum amount of human effort. Since it was deemed necessary to keep the parallel and alignable/aligned levels separate (see Section 3.2.2), two distinct analyses were carried out: (i) automatic segmentation accuracy vs. manual segmentation; and (ii) automatic alignment accuracy (see Section 3.2.2). Results are presented in Table 6, which compares for each range the average length variation rate, the average probability of correctly machine-*segmented* sentences, and the average probability of correctly machine-*aligned* sentences. The complete data set obtained from these analyses is presented in Appendix C.

LENGTH VARIATION RANGE	LENGTH VARIATION RATE [AVG]	SEGMENTATION ACCURACY [AVG]	ALIGNMENT ACCURACY [AVG]
-35% ≤ δ > -25%	-29.20206555	0.848621613	0.893113928
-25% ≤ δ > -15%	-20.86744282	0.875706663	0.892332554
-15% ≤ δ > -5%	-11.01013074	0.824349895	0.918786379
-5% ≤ δ ≥ +5%	-1.168811694	0.86162243	0.943869059
+5% < δ ≥ +15%	9.98585856	0.885306861	0.905709632
+15% < δ ≥ +25%	19.22908566	0.86202934	0.943921523
+25% < δ ≥ +35%	30.6796352	0.808919677	0.8399772

Table 6. Bitext automatic segmentation and alignment results for each range.

Results reveal that the overall automatic alignment accuracy is rather high (0.8923 to 0.9439), with the exception of bitexts in the +25%/+25% range (0.84). On the other hand, the overall automatic segmentation accuracy is quite low compared to the manual segmentation. Indeed, for all the length variation ranges values below 0.89 are obtained. This may be due to several factors, including the limitations of the software, the fine-tuning of the segmentation rules and the structure of the texts. However, the main reasons observed for segmentation (and alignment) errors in our sample are either missing punctuation or the presence of bibliographic reference lists in at least one of the bitext halves (see Appendix C, and more specifically, the NOTE column). In the former case, the aligner is not able to recognize sentence

ending strings, which leads to segmentation errors and misaligned bitext links. By way of example, Table 7 shows part of an automatically segmented/aligned bitext, where the majority of sentence links present misaligned material due to missing punctuation. In line with the recall-maximizing method described in Section 3.2.2, the only correctly aligned sentences are indeed the first source and target sentences in the table (i.e. "*Evoluzione … neuroscienze*" || "*Knowledges … neuroscience.*").

SOURCE TEXT [IT]	**TARGET TEXT [ELF]**
Evoluzione ed integrazione delle conoscenze: psicologia e neuroscienze Le teorie delle emozioni Le emozioni e il paradigma della complessità Strutture e funzioni cerebrali delle attività emotive Maturazione e sviluppo del sistema regolatore delle emozioni Psicopatologia delle emozioni Epidemiologia e problemi clinici Gli strumenti di valutazione delle funzioni emotive e cognitive I disturbi: diagnosi e presa in cura Il trattamento dei disturbi emozionali nuovi modelli di ricerca ~~~ ###Metodi	Knowledges evolution and integration: psychology and neuroscience.
lezioni frontali	Emotional theories.
incontri di approfondimento ~~~ ###Tipo	Emotions and the complex paradigm.
esame orale che mira a valutare il raggiungimento degli obiettivi didattici:	Brain structures and functions of emotional activities.
saper valutare i primi segni e sintomi dei disturbi emotivi	Maturation and development of emotions regulatory system.
saper discriminare tra diagnosi categoriale e diagnosi dimensionale	psychopathology of emotions Epidemiology and clinical problems.
conoscere le principali teorie sulle emozioni	Assessment tools of cognitive end emotional functions.
conoscere i principali strumenti di valutazione psicopatologica	Diagnosis and treatment of emotional disorders.
conoscere le principali tappe dello sviluppo emotivo	new research models ~~~ ###Metodi
conoscere la teoria della regolazione emotiva applicata alla psicopatologia	Frontal lectures ~~~ Deeping meeting
disturbi emotivi e modalità di trattamento	###Tipo ~~~ oral examination

Table 7. Example of segmented and aligned bitext with missing punctuation.

Similar results can be observed in the case of bibliographic reference lists in the bitext. In fact, name abbreviations in the bibliography often lead to misaligned links such as those shown in Table 8 below.

SOURCE TEXT [IT]	TARGET TEXT [ELF]
E.	E.
A. Albaugh, An autocrat's toolkit: adaptation and manipulation in 'democratic' Cameroon, in Democratization, 18, 2, 2011	A. Albaugh, An autocrat's toolkit: adaptation and manipulation in 'democratic' Cameroon, in Democratization, 18, 2, 2011
E.	E.
Green, Decentralization and political opposition in contemporary Africa: evidence from Sudan and Ethiopia, in Democratization, 18, 5, 2011	Green, Decentralization and political opposition in contemporary Africa: evidence from Sudan and Ethiopia, in Democratization, 18, 5, 2011
E.	E.
Hillbom, Botswana: a development-oriented gape-keeping state, in African Affairs, 111/442, 2012	
S.~~~ A. Bezabeh, Citizenship and the logic of sovereignty in Djibouti, in African Affairs, 110/441, 2011	Hillbom, Botswana: a development-oriented gape-keeping state, in African Affairs, 111/442, 2012 S. A. Bezabeh, Citizenship and the logic of sovereignty in Djibouti, in African Affairs, 110/441, 2011

Table 8. Random examples of automatically segmented and aligned bibliographic reference lists.

The author was unable to solve the problem of name abbreviations at the beginning of the sentence. On the contrary name abbreviations in the middle of the sentences (e.g. "*A. Albaugh*") are not segmented as a result of the fine-tuning of the software's segmentation rules (see Section 3.2.2). By way of example, the correct segmentation and alignment of the bibliographic reference item in the first two rows of Table 8 are presented in Table 9 below.

SOURCE TEXT [IT]	TARGET TEXT [ELF]
E. A. Albaugh, An autocrat's toolkit: adaptation and manipulation in 'democratic' Cameroon, in Democratization, 18, 2, 2011	E. A. Albaugh, An autocrat's toolkit: adaptation and manipulation in 'democratic' Cameroon, in Democratization, 18, 2, 2011

Table 9. Example of correctly segmented and aligned bibliographic reference item.

As mentioned in Section 3.2.2, it was hypothesized that the alignment accuracy depended on the quality of the bitext segmentation, i.e. the more *correct* the segmentation, the higher the alignment quality. Results presented so far do not provide support for this hypothesis (see Table 6). Segmentation accuracy is consistently lower than alignment accuracy, and the two data sets also show different trends. This difference is illustrated in Figure 2: on the horizontal axis are the average length variation rates for each range, whereas on the vertical axis are the data obtained from the analyses of the automatic segmentation (in purple) and alignment (in green).

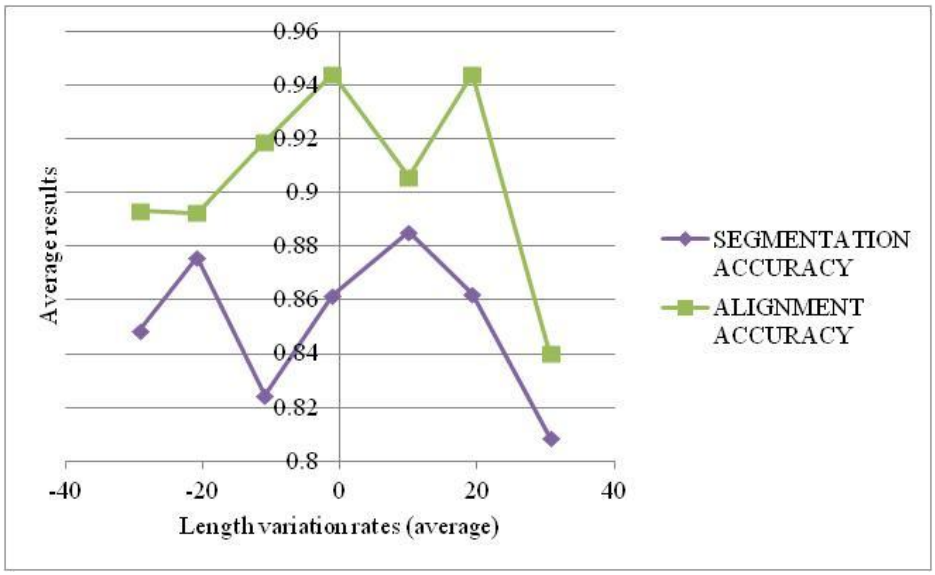

Figure 2. Bitext automatic segmentation and alignment trends.

It should be noted that alignment results follow an almost regular increasing trend in the negative ranges (see also Table 6). Indeed, alignment success rates tend to steadily increase up to the ±5% core interval (0.9439 correct sentence links), with the exception of a slight decrease from the -35%/-25% range (0.8931 accuracy) to the -25%/-15% range (0.8923 accuracy). On the contrary, segmentation accuracy does not follow a regular trend, showing often opposite results compared to the automatic alignment data in corresponding ranges. For instance, note the decreasing segmentation trend vs. the increasing alignment trends in the ranges represented in Figure 3 below.

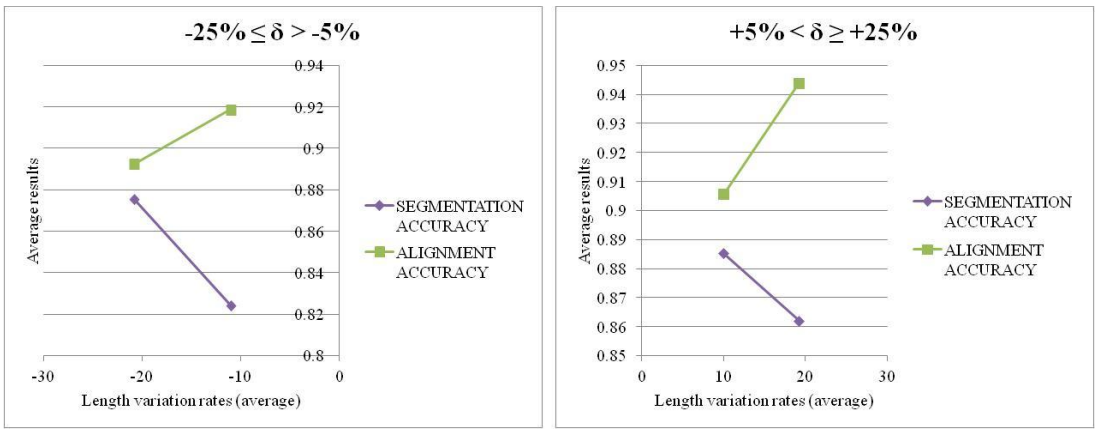

Figure 3. Bitext automatic segmentation vs. alignment in the -25%/-5% and the +5%/+25% ranges.

Likewise, the segmentation and alignment results in the range of -35%/-15% and -5%/+15% show opposite trends (see Figure 4 below).

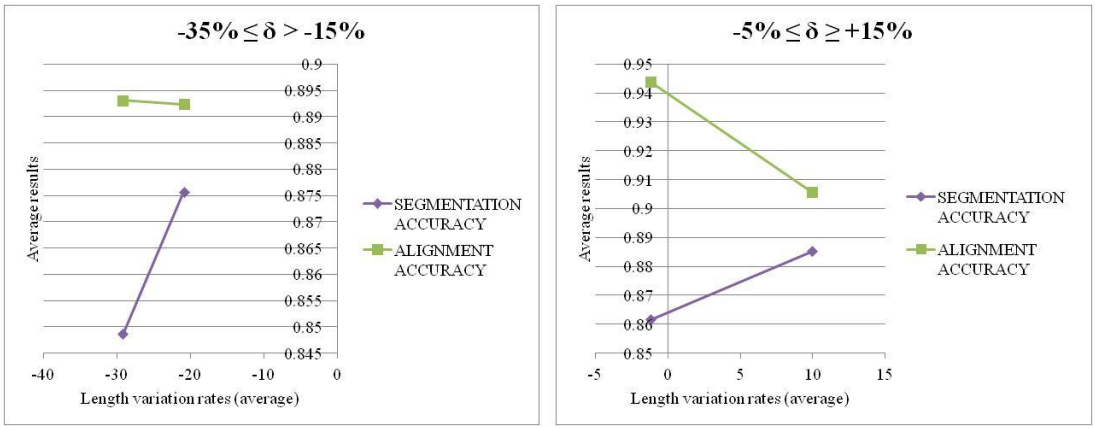

Figure 4. Bitext automatic segmentation vs. alignment in the -35%/-25% and -5%/+15% ranges.

This suggests that there is no direct connection between segmentation accuracy and alignment success rate. In fact, bitext alignment may yield, and it actually yields, high results even if the segmentation output is poor. The most striking example of this kind is the data set recorded for the -15%/-5% range, where 0.8243 segmentation accuracy corresponds to 0.9188 alignment accuracy (see Table 6 above). However, it should be noted that these results depend heavily on the notion of segmentation and alignment accuracy adopted in this study. Alignment accuracy is indeed defined in much looser terms than segmentation (see Section 3.2.2); therefore, results could vary to a great extent if one adopted a more restrictive definition of the alignment.

High-quality segmentation does not therefore necessarily lead to more accurate alignment results, and, conversely, wrong segmentation might not affect negatively the bitext alignment. Indeed, segmentation and alignment data from specific bitexts reveal that different results may be obtained. This is proven by the contrasting data recorded for the -25%/-15% range. In this respect, Table 10 lists in bold the most representative examples from said range (see Appendix C for complete data).

TOT_SENT	TOT_SENT_ SEGM	SEGMENTATION ACCURACY	TOT_SENT_ ALIGN	ALIGNMENT ACCURACY
33	33	**1**	33	**1**
24	24	**1**	22	**0.916666667**
110	106	**0.963636364**	93	**0.845454545**
65	45	**0.692307692**	54	**0.830769231**
48	37	**0.770833333**	48	**1**

Table 10. Bitext automatic segmentation vs. alignment: non-average data from the -25%/-15% range.

While alignment accuracy follows a regular increasing trend in negative ranges, positive ranges show a certain irregularity (see Table 6 and Figure 2). In fact, unlike the decreasing alignment trend in the positive ranges, the +15%/+25% range displays the highest alignment success rate in the sample, i.e. 0.9439. This is due to a practical reason. Noisiness in bitexts within positive and negative peripheral ranges (i.e. -35%/-15%; +15%/+35%) is often due to missing content/sections, hence their marked variation in characters. In most cases, however, existing source and target content within these ranges is highly parallel. Findings reveal that the aligner performs better on these types of bitexts than text pairs with different noisiness patterns (e.g. missing punctuation, bibliography, and so forth). Accordingly, Table 11 shows an example of 100% accurate alignment from the +15%/+25% range.

SOURCE TEXT [IT]	**TARGET TEXT [ELF]**
###Corso di laurea: MEDICINA E CHIRURGIA ###Titolo: EMATOPATOLOGIA C.I.	###Corso di laurea: MEDICINA E CHIRURGIA ###Titolo: EMATOPATOLOGIA C.I.
###Programma	###Programma
1) Test di clonalità in Ematologia	1) Clonality assays in hematology
2) Tecniche molecolari ad altra resa in onco-Ematologia	2) High-throughput tecnohnology in onco-hematology
3) Patologia molecolare del Linfoma di Hodgkin	3) Molecular pathology in Hodgkin Lymphoma
4) Patologia molecolare dei Linfomi non Hodgkin B	4) Molecular pathology in B-cell non Hodgkin Lymphomas
5) Patologia molecolare dei Linfomi non Hodgkin T	5) Molecular pathology in T-cell non Hodgkin Lymphomas
###Metodi	###Metodi
Didattica frontale	Face to face learning
Discussione di casi clinici	Clinical cases discussion
###Tipo	###Tipo
Quiz a risposta multipla	Multiple choice tests
###Obiettivi	###Obiettivi
	1) To understaind the molecular basis of the pathogenesis of hematopoietic tumors with special reference to malignant lymphomas;
	2) To develop a critical approach to the usage of molecular biomarkers as diagnostic, prognostic and therapy orientering tools
###Supporti	###Supporti
Diapositive	Slides

Table 11. Example of 100% accurate machine-aligned bitext in the +15%/+25% range.

The latter example may also prove useful to show that 100% alignment accuracy is achieved by several bitexts in *all* the length variation ranges (this is not the case of 100% bitext parallelism; see Section 4.1.1). Table 12 presents the distribution of 100% correctly aligned items in the sample (see also Appendix C).

LENGTH VARIATION RANGE	100% ALIGNMENT ACCURACY
-35% ≤ δ > -25%	5/15 (33.3%)
-25% ≤ δ > -15%	3/15 (20%)
-15% ≤ δ > -5%	5/15 (33.3%)
-5% ≤ δ ≥ +5%	1/15 (6.7%)
+5% < δ ≥ +15%	1/15 (6.7%)
+15% < δ ≥ +25%	3/15 (20%)
+25% < δ ≥ +35%	2/10 (20%)
TOT.	**20/100 (20%)**

Table 12. Distribution of bitexts that display 100% automatic alignment accuracy.

Despite the overall low segmentation accuracy, several bitexts in *all* the ranges also display 100% automatic segmentation accuracy (see Table 13 below). Although findings reveal that 100% segmentation accuracy may often correspond to 100% alignment accuracy, it is not necessarily the case that the two data sets are recorded for the same bitexts (see Appendix C). This is partly confirmed by the different number of 100% correctly segmented vs. aligned bitexts within most of the ranges (see Table 12 and 13).

LENGTH VARIATION RANGE	100% SEGMENTATION ACCURACY
-35% ≤ δ > -25%	2/15 (13.3%)
-25% ≤ δ > -15%	2/15 (13.3%)
-15% ≤ δ > -5%	4/15 (26.7%)
-5% ≤ δ ≥ +5%	2/15 (13.3%)
+5% < δ ≥ +15%	3/15 (20%)
+15% < δ ≥ +25%	3/15 (20%)
+25% < δ ≥ +35%	2/10 (20%)
TOT.	**18/100 (18%)**

Table 13. Distribution of bitexts that display 100% automatic segmentation accuracy.

The third prong of the approach also investigated the existence of a direct connection among the bitext variation in characters, the degree of bitext parallelism

(Section 4.1.1) and the alignment success rate (presented in this section). Findings of this analysis will be presented in Section 4.2.

4.1.3 Resource-oriented qualitative analysis

The last prong of the approach explored the amount of automatically aligned content sent to a translation memory (TM) that could be leveraged to assist Italian non-native authors/translators in producing ELF academic course descriptions. From an MT perspective, it also explored the amount of automatically aligned content within a parallel corpus that could be used to train a domain-restricted statistical MT system (see Section 3.2.3). To this end, an evaluation of the accuracy of the TM/parallel corpus was performed for each length variation range. For practical purposes, accuracy of both the TM and the parallel corpus will be referred to as *TM accuracy* and bitext links will be referred to as *translation units* (*TUs*). As mentioned in Section 3.2.3, TM accuracy differs from the accuracy parameters presented so far inasmuch as the analysis ignores *n*-to-zero and zero-to-*n* sentence correspondences.

Overall, results reveal that TM accuracy is high. Table 14 shows the average results for each range, whereas complete data are presented in Appendix D.

LENGTH VARIATION RANGE	TM ACCURACY [AVG]
$-35\% \leq \delta > -25\%$	0.899516229
$-25\% \leq \delta > -15\%$	0.909506087
$-15\% \leq \delta > -5\%$	0.934948353
$-5\% \leq \delta \geq +5\%$	0.957372426
$+5\% < \delta \geq +15\%$	0.92075507
$+15\% < \delta \geq +25\%$	0.959325397
$+25\% < \delta \geq +35\%$	0.839989177

Table 14. TM accuracy: average results for each range.

With the exception of the bitexts in the +25%/+35% length variation range, displaying 0.84 average percentage of leveragable content, the average values obtained through the count of correct TUs go from 0.8995 (in the -35%/-25% range) to 0.9593 (in the +15%/+25% range). Again, accuracy tends to increase as the bitext length variation rate approaches 0. However, the core interval (±5%) does not display the highest value, which is instead recorded for the +15%/+25% length variation range. As also mentioned in Section 4.1.2, this is due to the good

performance of the aligner when faced with bitexts containing several missing sections, such as those in the -35%/-15% and +15%/+35% peripheral ranges.

Significantly, *all* the ranges include several bitexts displaying 100% TM accuracy. The distribution of 100% accurate TMs across the various ranges is presented in Table 15. The greatest number of 100% accurate TMs is found in the -15%/+5% range (in bold).

LENGTH VARIATION RANGE	100% ACCURATE TMs
$-35\% \leq \delta > -25\%$	5/15 (33.3%)
$-25\% \leq \delta > -15\%$	3/15 (20%)
$-15\% \leq \delta > -5\%$	**8/15 (53.3%)**
$-5\% \leq \delta \geq +5\%$	**6/15 (40%)**
$+5\% < \delta \geq +15\%$	2/15 (13.3%)
$+15\% < \delta \geq +25\%$	5/15 (33.3%)
$+25\% < \delta \geq +35\%$	2/10 (20%)
TOT.	**31/100 (31%)**

Table 15. Distribution of 100% accurate TMs in the sample.

An example of 100% accurate TM is provided in Table 16: grey-shaded rows are automatically discarded by the software in the process of creation of the TM.

SOURCE TEXT [IT]	TARGET TEXT [ELF]
###Corso di laurea: ARCHITETTURA ###Titolo: AFPG - LABORATORIO DI DIAGNOSTICA STRUTTURALE II	###Corso di laurea: ARCHITETTURA ###Titolo: AFPG - LABORATORIO DI DIAGNOSTICA STRUTTURALE II
###Programma	###Programma
L'attività che si intende svolgere attraverso questo laboratorio è incentrata principalmente sullo studio del comportamento meccanico dei materiali da costruzione, quali calcestruzzo, acciaio, cemento armato, legno e murature.	The activities to be carried out by this laboratory is focused mainly on the study of the mechanical behavior of construction materials, such as concrete, steel, reinforced concrete, wood and masonry.
Per determinare le caratteristiche meccaniche dei materiali è necessario condurre un'opportuna sperimentazione sui materiali stessi così come richiesto e descritto dalla normativa vigente.	In order to determine the mechanical characteristics of materials, it is necessary to conduct appropriate testing program as required and described by technical standards.
Tali prove saranno condotte presso un Laboratorio di certificazione dei materiali riconosciuto a livello internazionale con l'ausilio di tecnici specializzati.	These tests will be conducted only by internationally recognized certified materials laboratory with the help of expert technicians.
Conoscere i materiali e il loro comportamento è alla base del saper progettare in maniera consapevole.	The knowledge of the materials and their behavior represent the basis of being able to project in a conscious way.

Tale conoscenza diviene ancor più necessaria soprattutto quando si tratta di interventi di recupero di edifici già esistenti.	Such knowledge becomes even more necessary especially when it comes to recovery interventions of existing buildings.
###Metodi	###Metodi
Lezioni frontali, utilizzo delle attrezzature nel Laboratorio LADS e visite a Laboratori Ufficiali	Lectures, use of equipment in the Laboratory LADS and visits to Official Laboratories
###Tipo	###Tipo
Esame della tesina sulle attività svolte, prova orale sulle conoscenze acquisite	Examination of the paper on the activities, oral examination on the knowledge acquired
###Obiettivi	###Obiettivi
L'attività che si intende svolgere presso questo laboratorio completa lo studio del comportamento meccanico dei materiali da costruzione, quali calcestruzzo, acciaio, cemento armato, legno e murature.	
Per determinare le caratteristiche meccaniche dei materiali è necessario condurre un'opportuna sperimentazione mediante prove distruttive e microdistruttive sui materiali stessi, così come richiesto e descritto dalla normativa vigente.	
Lo studente dovrà acquisire la necessaria conoscenza delle operazioni connesse alla esecuzione delle prove distruttive che potranno essere eseguite presso un Laboratorio di certificazione dei materiali riconosciuto a livello nazionale	
###Supporti	###Supporti

Table 16. Example of 100% accurate TM/parallel corpus in the -35%/-25% range.

4.2 Discussion

Results presented so far will be now discussed and compared with the aim of answering the questions addressed by the present study, reported in the introductory section of Chapter 3. Data will be explored both horizontally and vertically, i.e. the whole data set obtained for each analysis will be compared to the data sets from the other analyses at *range* and *sample* level respectively.

4.2.1 Document length similarity vs. bitext parallelism

The evaluation of the extent of parallelism within a set of noisy bitexts (i.e. question 1 in Chapter 3) has already been discussed in Section 4.1.1. Results revealed that bitext parallelism in the sample follows a parabolic trend from the lowest length variation range to the highest range. According to this trend, it is likely to find the highest probability of translation equivalence in the ±5% core interval (i.e. 0.9576). On the other hand, the analysis also investigated the relation between bitext parallelism in the sample and document similarity (i.e. question 2). In Section 3.1, a close relation between document length similarity (characters) and translation equivalence has been hypothesized. Findings support this hypothesis, confirming that there actually exists a direct connection between them. This connection is clearly exemplified by the similar patterns of the two curves in Figure 5, where the blue curve shows the average length variation rates for each range and the red curve represents the average bitext parallelism data.

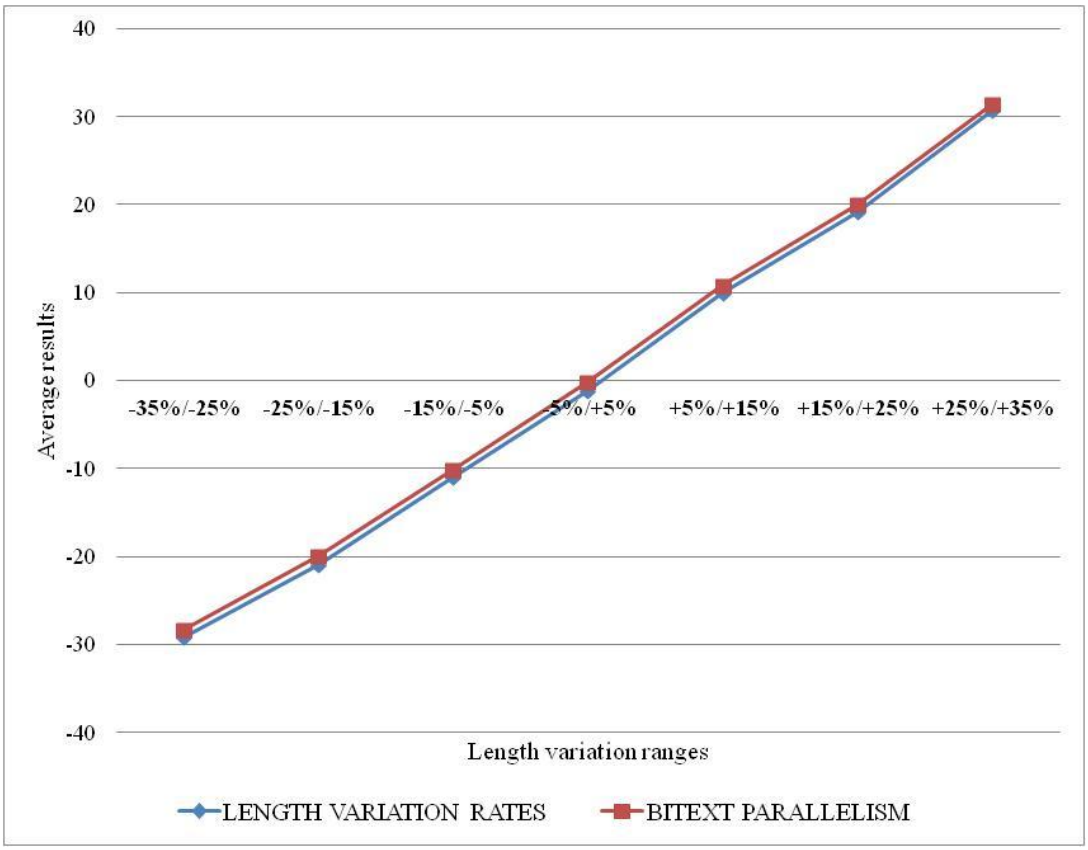

Figure 5. Length variation rates vs. bitext parallelism: similar trends.

These trends also confirm another hypothesis formulated in Section 3.1. Findings prove indeed that text pairs with negative values (i.e. those where Italian

texts tend to be longer than their English equivalents) are likely to be more parallel than their positive counterparts. Specifically, our hypothesis implied that bitexts displaying negative values in the range of (at least) 0%/-15% would yield the highest degree of parallelism among the ranges in the sample. On the one hand, this hypothesis is supported by the data reported in Table 1 (Section 4.1.1): bitexts in the range of -15%/-5% and ±5% display nearly the same (high) degree of parallelism (0.9547 and 0.9576 respectively). On the other hand, results exceeded our expectations, since high probability values are also recorded for lower negative ranges. In this respect, Table 17 compares the different degrees of parallelism recorded for symmetrical positive and negative ranges. Note that parallelism probabilities are consistently higher in negative ranges (left-hand side of the table) than in the corresponding positive ranges (right-hand side of the table); e.g., the 0.9003 parallelism probability in the -25%/-15% range is higher than the 0.8979 parallelism obtained for both the +5%/+15% range and the +15%/+25% ranges.

LENGTH VARIATION RANGE [NEGATIVE]	BITEXT PARALLELISM [AVG]	LENGTH VARIATION RANGE [POSITIVE]	BITEXT PARALLELISM [AVG]
-35% ≤ δ > -25%	0.850598909	**+25% < δ ≥ +35%**	0.720750859
-25% ≤ δ > -15%	0.900272415	**+15% < δ ≥ +25%**	0.897941311
-15% ≤ δ > -5%	0.954722531	**+5% < δ ≥ +15%**	0.897957767

Table 17. Bitext parallelism: comparison of average results from symmetrical ranges.

These results strengthen the hypothesis of overall higher translation equivalence probability between longer Italian texts and shorter equivalent English texts (see Section 3.1). In other words, it is more likely to find parallel bitexts among negative length variation ranges, i.e. where presumed English target texts have fewer characters than presumed Italian source texts.

As can be noticed, the core interval is not listed in Table 17. This is because the analysis of the ±5% range included both negative and positive values. Consequently, it might be objected that since the average result computed for the ±5% range also includes data obtained from text pairs with positive variation rates, the above-mentioned hypothesis is only partly confirmed. While this is a valid objection, it must be acknowledged that similar trends to the ones outlined so far can be observed for average values from the negative and positive sub-sets within the range under consideration (see Table 18 below). As can be noticed, bitexts with

negative variation rates in the range of -5%/0% (in the left-hand side of the table) tend to present higher probabilities of parallelism (i.e. 0.9718) than bitexts with corresponding positive values (i.e. 0.9327). While this might not always be the case for bitexts displaying symmetrical length variation rates in lower/higher ranges (i.e. less than -5% or greater than +5%), it is nonetheless a significant result to be taken into account for the creation of a reference resource for non-native authors/translators (e.g. parallel corpus, translation memory).

$-5\% \leq \delta \geq +5\%$			
LENGTH VARIATION RANGE [NEGATIVE]	**BITEXT PARALLELISM [AVG]**	**LENGTH VARIATION RANGE [POSITIVE]**	**BITEXT PARALLELISM [AVG]**
$-5\% \leq \delta > 0\%$	0.971792901	**$0\% < \delta \geq +5\%$**	0.932708164

Table 18. Bitext parallelism in symmetrical length variation rates within the ±5% range.

Another aspect to be taken into account in the latter respect is the 21% probability of finding ELF-ELF bitexts in the corpus, mentioned in Section 4.1.1. Performing language detection before automatic alignment is crucial in order to avoid non-negligible consequences on the quality of the parallel corpus and the resulting translation memory.

4.2.2 Document length similarity vs. bitext parallelism vs. bitext alignment accuracy

Question 3 addressed the extent of alignment accuracy of the bitexts in the sample. Results presented in Section 4.1.2 revealed that the number of correctly aligned sentences in the various ranges is high (0.8923 to 0.9439), with the exception of bitexts in the +25%/+35% range (0.84). Given the noisiness of the bitexts and considering that they have not been pre-processed, i.e. the only human intervention performed was the fine-tuning of the segmentation rules (see Section 3.2.2), the latter results are satisfactory for our purposes.

This aspect of the approach also explored the relation between document length similarity, bitext parallelism and alignment accuracy (i.e. question 4). Findings reveal that the average results of these three analyses are actually closely related to each other. The strong connection between document similarity and bitext parallelism has already been confirmed in Section 4.2.1 (see Figure 5 in particular).

Likewise, bitext parallelism and alignment accuracy show similar results. Indeed, as in the case of bitext parallelism, alignment accuracy tends to increase as the length variation decreases (i.e. approaches 0), peaking in the core interval (±5%). As mentioned in Section 4.1.2, very similar values were recorded for the core interval and the +15%/+25% range (i.e. 0.9439): this contrasts sharply with the steadily decreasing trend that characterizes positive length variation rates in the analysis of bitext parallelism. Figure 6 exemplifies what has been discussed so far: the average length variation rates for each range are displayed along the horizontal axis, whereas the average bitext parallelism (in red) and alignment success rate (in green) are shown on the vertical axis.

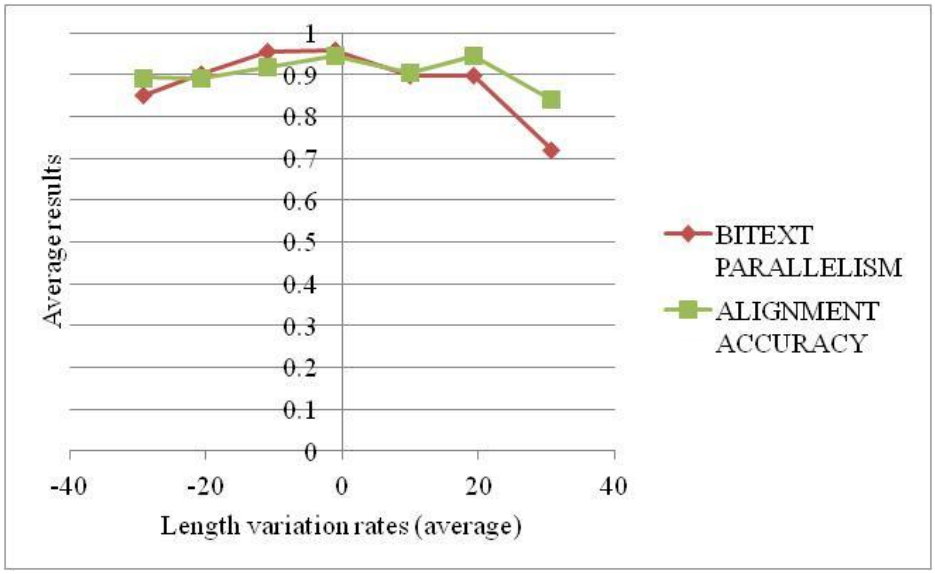

Figure 6. Bitext parallelism and automatic alignment trends.

This close relation is also confirmed at range level by the slight difference between bitext parallelism and alignment accuracy. Said difference ranges from -0.0079 to +0.1192 (see the last column of Table 19 below). In this case, negative values imply that alignment results are lower than data on bitext parallelism. Accordingly, positive differences imply that alignment results are higher than bitext parallelism results. While alignment accuracy presupposes bitext parallelism, it also examines *n*-to-zero and zero-to-*n* sentence correspondences (unlike bitext parallelism). This is the reason why alignment accuracy is sometimes higher than the probability of translation equivalence within the bitexts.

LENGTH VARIATION RANGE	BITEXT PARALLELISM [AVG]	ALIGNMENT ACCURACY [AVG]	DIFFERENCE
-35% ≤ δ > -25%	0.850598909	0.893113928	**0.042515**
-25% ≤ δ > -15%	0.900272415	0.892332554	**-0.00794**
-15% ≤ δ > -5%	0.954722531	0.918786379	**-0.03594**
-5% ≤ δ ≥ +5%	0.957580269	0.943869059	**-0.01371**
+5% < δ ≥ +15%	0.897957767	0.905709632	**0.007752**
+15% < δ ≥ +25%	0.897941311	0.943921523	**0.04598**
+25% < δ ≥ +35%	0.720750859	0.8399772	**0.119226**

Table 19. Difference between bitext parallelism and automatic alignment accuracy at range level.

As can be noticed in Table 19, the largest difference is observed for the +25%/+35% range (i.e. 0.1192). If we excluded said range because of its low probability of parallelism, values would go from -0.0079 to +0.046. This leads us to believe that there exists a strong connection between 85% or higher bitext parallelism rate and corresponding bitext alignment success rate. This confirms and expands the sample-level answer given above to the first part of question 4, i.e. whether any connection can be observed between the alignment success rate and the degree of parallelism of the aligned text pairs (see Figure 6).

The second part of question 4 concerned the existence of any relation between alignment accuracy and document length similarity. As mentioned in Section 4.2.1, document length similarity is closely related to bitext parallelism. On the other hand, the strong connection between bitext parallelism and alignment accuracy has just been confirmed (see Figure 6 and Table 19). A close relation between document length similarity and alignment accuracy might consequently be hypothesized. The increasing and decreasing trends that characterize length variation rates and alignment results in symmetrical negative and positive ranges confirm this hypothesis (see Table 20).

LENGTH VARIATION RANGE	LENGTH VARIATION RATE [AVG]	BITEXT PARALLELISM [AVG]	ALIGNMENT ACCURACY [AVG]
-35% ≤ δ > -25%	-29.20206555	0.850598909	0.893113928
-25% ≤ δ > -15%	-20.86744282	0.900272415	0.892332554
-15% ≤ δ > -5%	-11.01013074	0.954722531	0.918786379
-5% ≤ δ ≥ +5%	-1.168811694	0.957580269	0.943869059
+5% < δ ≥ +15%	9.98585856	0.897957767	0.905709632
+15% < δ ≥ +25%	19.22908566	0.897941311	0.943921523
+25% < δ ≥ +35%	30.6796352	0.720750859	0.8399772

Table 20. Length variation rates vs. bitext parallelism vs. automatic alignment.

The patterns of the three curves in Figure 7 prove once more that document similarity, bitext parallelism and automatic alignment are characterized by very similar trends.

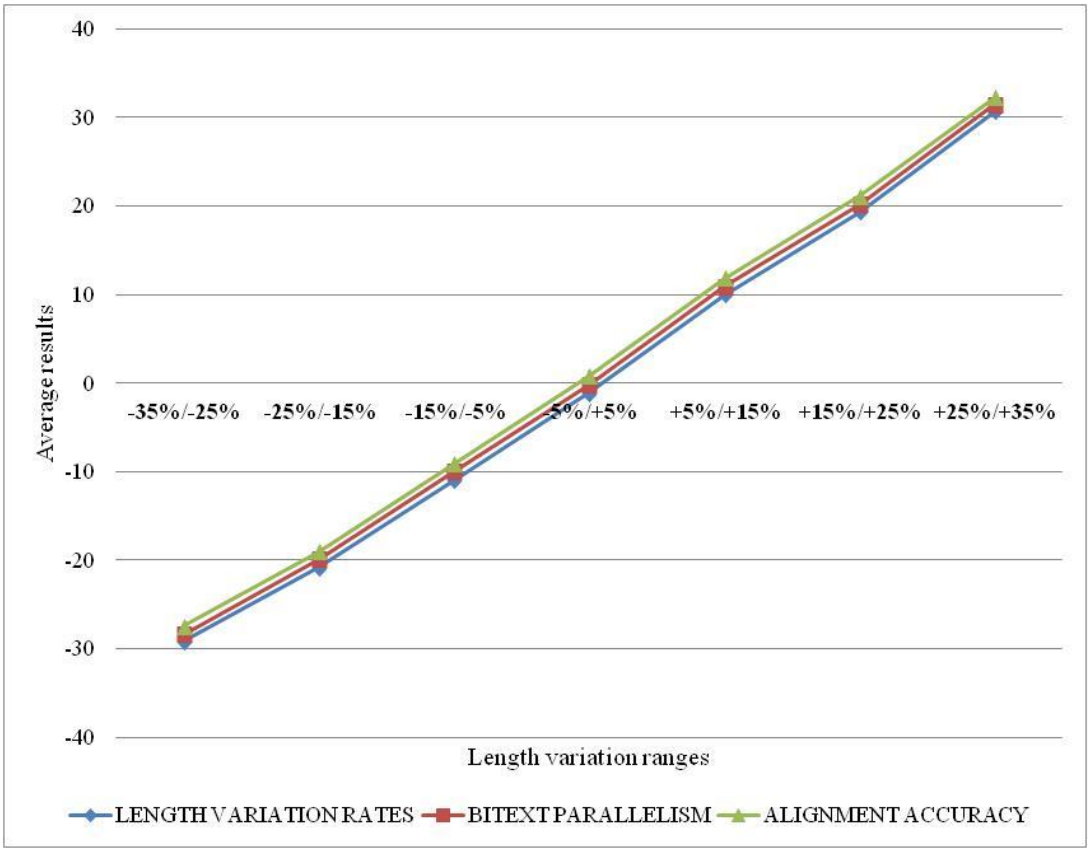

Figure 7. Document length similarity vs. bitext parallelism vs. automatic alignment: similar trends.

4.2.3 Document length similarity vs. bitext parallelism vs. bitext alignment accuracy vs. resource leverageability

The last question of the study investigated the amount of content that could be leveraged if we were to create a parallel corpus and/or a translation memory. As mentioned in Section 4.1.3 both of them will be referred to as TM. Section 4.1.3 revealed that a large amount of content can be leveraged through automatic alignment (i.e. 0.8995 to 0.9593), except for bitexts in the +25%/+35% range (i.e. 0.84). Findings also revealed that TM accuracy tends to increase in the negative ranges and decrease in the positive ranges, with the exception of the +15%/+25% range. Similarly to the automatic segmentation and alignment results, all the ranges include 100% accurate TMs (see Section 4.1.3). However, the number of these instances is higher than the 100% cases recorded in the former analyses, i.e. 31% vs.

20% (alignment) and 18% (segmentation) (see Sections 4.1.1 and 4.1.2). Moreover, unlike the segmentation vs. alignment analysis, it is highly probable for a 100% accurate alignment to return a 100% accurate TM. When this is not the case, alignment success rates leading to 100% TM accuracy are still considerably high, i.e. not lower than 0.9454 (see Appendix D). Consequently, 100% TM accuracy always and exclusively corresponds to 100% or slightly lower alignment accuracy in our sample. It should be recalled that this is not necessarily the case with segmentation vs. alignment accuracy, which are not related to each other. In fact, extremely low segmentation accuracy may sometimes lead to 100% correctly aligned bitexts (see Section 4.1.2).

Findings also reveal that TM and alignment accuracy show equivalent patterns. In fact, on average, the percentage of *correct* TUs for each range mirrors the percentage of correctly aligned sentences for each range. In this respect, Table 21 compares the two data sets under consideration.

LENGTH VARIATION RANGE	ALIGNMENT ACCURACY [AVG]	TM ACCURACY [AVG]
-35% ≤ δ > -25%	0.893113928	0.899516229
-25% ≤ δ > -15%	0.892332554	0.909506087
-15% ≤ δ > -5%	0.918786379	0.934948353
-5% ≤ δ ≥ +5%	0.943869059	0.957372426
+5% < δ ≥ +15%	0.905709632	0.92075507
+15% < δ ≥ +25%	0.943921523	0.959325397
+25% < δ ≥ +35%	0.8399772	0.839989177

Table 21. Bitext automatic alignment vs. TM accuracy for each range.

Significantly, all the ranges in Table 21 display extremely close alignment vs. TM results. Figure 8 exemplifies the relation between automatic alignment and TM accuracy. Similarly to the previous analyses, the average length variation rates for each range are displayed on the horizontal axis, whereas the alignment average results (in green) and the TM average accuracy (in orange) are shown on the vertical axis.

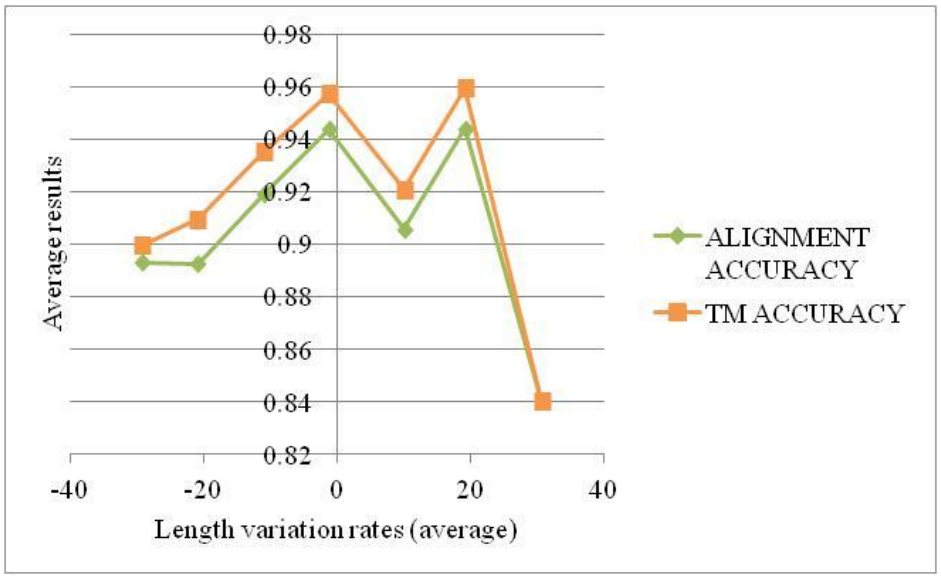

Figure 8. Bitext automatic alignment and TM accuracy trends.

Figure 8 clearly illustrates that there exist a close and direct connection between automatic alignment and TM accuracy. First of all, as in the case of bitext alignment, the highest TM accuracy was recorded for the +15%/+25% range (i.e. 0.9593), closely followed by the ±5% core interval (i.e. 0.9574). Secondly, negative ranges are characterized by increasing TM accuracy, whereas positive ranges tend to display decreasing values. The +15%/+25% range is once more an exception to said decreasing pattern (see also Section 4.1.2). There is only one contrasting trend between the two curves, i.e. the decreasing alignment success rate vs. the increasing TM accuracy in the -35%/-15% range (see also Table 22).

The close relation between automatic alignment success rate and TM accuracy is also confirmed at range level. In this case, the difference between the results of the two analyses ranges from +0.00001 to +0.0172 (see Table 22 below). This means that TM accuracy is always higher than automatic alignment accuracy in our sample.

LENGTH VARIATION RANGE	ALIGNMENT ACCURACY [AVG]	TM ACCURACY [AVG]	DIFFERENCE
-35% ≤ δ > -25%	0.893113928	0.899516229	**0.006402**
-25% ≤ δ > -15%	0.892332554	0.909506087	**0.017174**
-15% ≤ δ > -5%	0.918786379	0.934948353	**0.016162**
-5% ≤ δ ≥ +5%	0.943869059	0.957372426	**0.013503**
+5% < δ ≥ +15%	0.905709632	0.92075507	**0.015045**
+15% < δ ≥ +25%	0.943921523	0.959325397	**0.015404**
+25% < δ ≥ +35%	0.8399772	0.839989177	**0.000011977**

Table 22. Difference between automatic alignment and TM accuracy at range level.

Likewise, the relation between bitext parallelism and TM accuracy is determined through the computation of the difference between the average results of the two analyses at range level. In this respect, Table 23 shows a variation span ranging from -0.0002 to +0.1192. Similarly to the data in Section 4.2.2, negative values imply that bitext parallelism is higher than TM accuracy, and vice versa. Unlike the comparison between bitext parallelism and alignment accuracy made in Section 4.2.2, both bitext parallelism and TM accuracy ignore *n*-to-zero and zero-to-*n* correspondences. On the other hand, as in the case of alignment accuracy, TM accuracy presupposes bitext parallelism. Thus, it might seem strange that TM accuracy presents higher data than bitext parallelism. In this case, this is due to the different object of analysis, i.e. translation units vs. sentences respectively. Data on bitext parallelism at sentence level are therefore not reflected in TM accuracy results, where the parallelism of the aligned content is evaluated differently from the former analysis (see Section 3.2.3).

LENGTH VARIATION RANGE	BITEXT PARALLELISM [AVG]	TM ACCURACY [AVG]	DIFFERENCE
$-35\% \leq \delta > -25\%$	0.850598909	0.899516229	**0.048917**
$-25\% \leq \delta > -15\%$	0.900272415	0.909506087	**0.009234**
$-15\% \leq \delta > -5\%$	0.954722531	0.934948353	**-0.01977**
$-5\% \leq \delta \geq +5\%$	0.957580269	0.957372426	**-0.00021**
$+5\% < \delta \geq +15\%$	0.897957767	0.92075507	**0.022797**
$+15\% < \delta \geq +25\%$	0.897941311	0.959325397	**0.061384**
$+25\% < \delta \geq +35\%$	0.720750859	0.839989177	**0.119238**

Table 23. Difference between bitext parallelism and TM accuracy at range level.

Data in Table 23 reveal that there exists a close relation between 85% or higher bitext parallelism rate and TM accuracy. Again, the exclusion of the +25%/+35% range (i.e. +0.1192 difference) for its high degree of noisiness leads to a difference span ranging from -0.0002 to +0.0614. This confirms the strong range-level and sample-level connection between the results of the two analyses.

Last but not least, the increasing and decreasing trends that characterize length variation rates and the results in terms of leverageable content in symmetrical negative and positive ranges prove that there also exists a direct relation between document length similarity and TM accuracy. Consequently, it can be said that all the aspects examined in the present study are closely related to each other. This is confirmed by the average results of *all* the analyses discussed so far (except for the

analysis of the automatic segmentation), compared in Table 24 and illustrated in Figure 9.

LENGTH VARIATION RANGE	LENGTH VARIATION RATE [AVG]	BITEXT PARALLELISM [AVG]	ALIGNMENT ACCURACY [AVG]	TM ACCURACY [AVG]
-35% ≤ δ > -25%	-29.20206555	0.850598909	0.893113928	0.899516229
-25% ≤ δ > -15%	-20.86744282	0.900272415	0.892332554	0.909506087
-15% ≤ δ > -5%	-11.01013074	0.954722531	0.918786379	0.934948353
-5% ≤ δ ≥ +5%	-1.168811694	0.957580269	0.943869059	0.957372426
+5% < δ ≥ +15%	9.98585856	0.897957767	0.905709632	0.92075507
+15% < δ ≥ +25%	19.22908566	0.897941311	0.943921523	0.959325397
+25% < δ ≥ +35%	30.6796352	0.720750859	0.8399772	0.839989177

Table 24. Length variation rates vs. bitext parallelism vs. automatic alignment vs. TM accuracy.

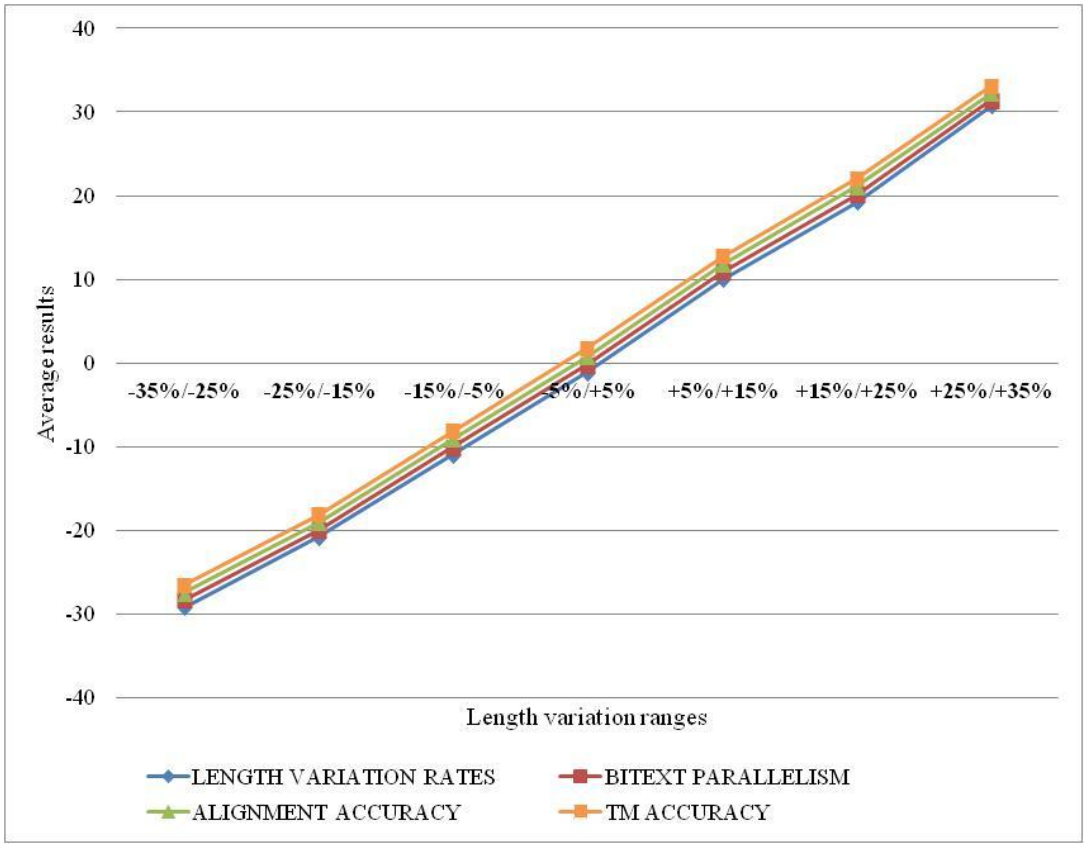

Figure 9. Document length similarity vs. bitext parallelism vs. automatic alignment vs. TM accuracy: similar trends.

Ignoring the CODE-UniBO corpus categorization according to length variation classes (see Section 3.1), similar trends are also observed among the average results of all the bitexts in the sample. In this respect, Table 25 compares the average extent of parallelism, alignment accuracy and percentage of leverageable

content (or TM accuracy) in all the Italian-ELF human-produced bitexts in the sample.

BITEXT PARALLELISM [AVG_SAMPLE]	ALIGNMENT ACCURACY [AVG_SAMPLE]	DIFFERENCE
0.892685685	0.906312792	**0.013627**
BITEXT PARALLELISM [AVG_SAMPLE]	**TM ACCURACY [AVG_SAMPLE]**	
0.892685685	0.918758312	**0.026073**
ALIGNMENT ACCURACY [AVG_SAMPLE]	**TM ACCURACY [AVG_SAMPLE]**	
0.906312792	0.918758312	**0.012446**

Table 25. Sample parallelism vs. automatic alignment accuracy vs. TM accuracy.

The close relation between these three aspects is once more confirmed by: (i) the +0.0136 difference between sample alignment success rate and extent of parallelism; (ii) the +0.0261 difference between TM accuracy and extent of parallelism in the sample; and (iii) the +0.0124 difference between TM accuracy and sample alignment success rate.

Although results in Table 25 are sufficiently good considering the minimum amount of human effort required, they could be improved by aligning only bitexts that would return the most profitable results. This aspect will be discussed in greater detail in Section 4.4.

4.3 Follow-up: creating an institutional academic resource for non-native authors and translators

4.3.1 Bitext selection

In the light of the data presented in Section 4.1 and discussed in Section 4.2, we decided to create a reference resource for non-native authors/translators through the automatic alignment of 85% (or higher) parallel bitexts in the CODE-UniBO corpus. In fact, findings reveal that it is more likely for these items to yield high accuracy (i.e. equal or greater than 0.85) in terms of both alignment and leverageable content (see Sections 4.2.2 and 4.2.3). The majority of the length variation ranges examined in the previous sections would give acceptable results in this respect. The only exception is the +25%/+35% span, which yielded 0.84 results for both alignment and TM accuracy. Consequently, bitexts that displayed length differences in the

range of -35% to +25% were automatically aligned, whereas bitexts that displayed lower or greater length variation rates were ignored. This decision is justified by the fact that data presented in Table 25 (Section 4.2.4) slightly improve if we exclude the +25%/+35% range from the sample: +0.0169 increase in parallelism, +0.0065 increase in alignment success rate, and +0.0078 increase in TM accuracy. Table 26 compares the data obtained from bitexts in all the ranges (row 1; see also Table 25) with the data obtained from the bitexts in the range of -35%/+25% (in bold in row 2). The increase in each analysis is reported in the last row of the table.

	BITEXT PARALLELISM [AVG SAMPLE]	ALIGNMENT ACCURACY [AVG SAMPLE]	TM ACCURACY [AVG SAMPLE]
1 [including the +25%/+35% range]	0.892685685	0.906312792	0.918758312
2 [excluding the +25%/+35% range]	**0.909637006**	**0.91285292**	**0.926524282**
INCREASE	0.016951321	0.006540128	0.00776597

Table 26. Sample parallelism vs. automatic alignment accuracy vs. TM accuracy in the -35%/+25% range.

4.3.2 Parallel corpus and TM creation

The total number of bitexts displaying a length variation rate in the range of -35%/+25% is 3,146 out of the 3,263 bitexts initially extracted from the CODE-UniBO corpus (cf. Chapter 2). In order to create a reference resource for non-native authors and translators, the pairs of English source and target texts and Italian source and target texts were semi-automatically excluded from the set of bitexts examined in this study. The automatically translated bitext in the sample was also excluded (see also Section 4.1.1). However, if the corpus contained other similar bitexts, they might have been included.

In practical terms, a total number of 3,044 bitexts were therefore used to build a parallel corpus of Italian-ELF academic course descriptions through automatic alignment, i.e. the CODE-UniBO-Par corpus. The parallel corpus contains

1,269,072 Italian tokens (i.e. words) and 1,123,970 English tokens,[25] and it is composed of 99,181 bitext links containing several bitext link correspondences other than *n*-to-zero and zero-to-*n* sentence correspondences.

The CODE-UniBO-Par corpus was eventually exported to TMX format. The resulting TM contains 99,181 TUs.

4.4 The contribution of the present study and directions for future research

In what follows, we illustrate the relevance of the findings discussed so far for translation research, the main limitations of this study as well as some suggestions for future research. First of all, results suggest that it is possible – and indeed profitable – to use length similarity (in terms of characters) as a clue to establish the probability of parallelism of a pair of texts given as translation equivalents. Results also support the hypothesis that it is more likely to record a high probability of translation equivalence for bitexts where the Italian source text is longer than the English target text and that this probability reaches its peak in extremely similar bitexts in terms of number of characters (i.e. -15%/0%) (see Section 4.2.1). Secondly, findings point at a direct relation between document length similarity, bitext parallelism, alignment success rate and TM accuracy. Results show that these four aspects follow very similar trends: with few exceptions, data show increasing trends in the negative ranges (i.e. bitexts containing longer Italian texts) and decreasing trends in the positive ranges (i.e. bitexts containing longer English texts). This means that parallelism, alignment and TM accuracy tend to be higher in the negative ranges as the bitext length variation rates approach 0. Few exceptions in terms of alignment and TM accuracy concern bitexts with considerable amount of missing information in the +15%/+25% positive peripheral range. Finally, range-level analysis revealed that 85% (or more) parallel bitexts are most likely to achieve high-quality alignment and leverageability results (i.e. greater than or equal to 85%) with the minimum amount of human effort (see Sections 4.2.2 and 4.2.3). This ultimately hints at the possibility that alignment accuracy and resource

[25] The count was performed using the "wc" Unix command-line utility. The count also includes metadata (e.g. ###*Programma*; cf. Section 1.4.2), which are considered as one token each.

leverageablity can be related to the difference in terms of number of characters (i.e. document length similarity) between a pair of Italian and English texts.

One of the limits of the comparative analyses conducted in the present study is that they only examine a specific genre (i.e. institutional academic genre) and text type (i.e. web-based academic course descriptions) belonging to a single academic institution (i.e. the University of Bologna). The heuristic approach based on character, sentence and translation unit comparison (presented in Chapter 3) may be used for future comparative investigations on similar texts produced by different Italian academic institutions. Building on the approach adopted in this study, different types of Italian-EFL bitexts in the institutional academic genre as well as different types of texts belonging to domains other than higher education may also be examined. Results of these analyses may be compared to those obtained in our study in order to assess and generalize the validity of the argument presented here, i.e. that there exists a direct connection between document similarity, bitext parallelism, alignment accuracy and resource leverageability in academic course descriptions produced by Italian universities, and that the same connection also underlies different types of Italian-ELF bitexts produced by Italian authors and/or translators in different domains.

Similar ELF comparative studies may investigate this connection in academic course descriptions and other bitexts produced by several European academic institutions. Studies are indeed required which compare our findings with those obtained for pairs of English and another Romance language, like e.g. Spanish or French texts. It might be hypothesized that the close relatedness of the latter languages with the Italian language would lead to similar results. If confirmed, a common hypothesis for Romance languages may be developed which generalizes and extends the results presented in this contribution to several domains and text types. On the other hand, it would be interesting to also compare the results of this study with results from similar studies on ELF and another Germanic language, like e.g. German. Similarly to Italian, it is common knowledge that German texts tend to display a greater number of characters than English texts. However, the nature of German morphology and syntax leads us to hypothesize that the variation in characters between German and English texts would be higher than in the case of Italian-ELF bitexts. In the case of German-ELF bitexts, the corpus categorization may include different length variation ranges from those presented in this study, or

be carried out in terms of variation in words rather than variation in characters. Results may eventually be compared with data obtained from the various analyses in this study, establishing possible relations between the Italian-ELF (as well as Spanish/French-ELF) and the German-ELF trends. Common trends between these (and additional) language pairs would lay the ground for a more comprehensive theory at European level.

Comparative studies like the one presented here may also be useful from a resource-oriented perspective. Indeed, the original contribution made by the study provides useful insights into the specific bitexts that should be used for the creation of a reference resource to assist Italian non-native authors and/or translators working with institutional academic English. In this respect, Ferraresi and Bernardini (forthcoming) suggest indeed two lines of research to further public/private institutional and administrative interventions to increase bilingualism in the European academic environment. One of them is "the implementation of tools for assisting non-native writers in producing appropriate texts in [institutional academic English]" (Ferraresi and Bernardini, forthcoming) (see also Section 1.2). To the best of our knowledge, the only existing resource of this type at the time of writing is the BTS machine translation service (Depraetere et al. 2011). The service, however, does not include the Italian-English language pair (see Section 1.3.1). Data presented in this contribution may be a first attempt at providing useful information to extend BTS technology to the Italian language.

It might be objected that MT is still not the solution to the increasing demand for translations in the institutional academic domain because of the poor linguistic quality of English content on university websites (Bernardini, 2014). Against this background, the most profitable bitexts in our study in terms of alignment accuracy and leverageability have been used to build a parallel corpus (i.e. the CODE-UniBO-Par corpus) and a translation memory of Italian-ELF web-based academic course descriptions (cf. Section 4.3). From a descriptive perspective, the CODE-UniBO-Par corpus may be used by corpus linguists to examine the ELF linguistic and translation choices made by non-native authors/translators and to set quality language and translation standards at national (i.e. Italian) and European level. From a practical perspective, the corpus may be used by Italian non-native authors as a reference resource for the production of ELF content on university websites. On the other hand, it may be used to train a high-quality, domain-restricted statistical MT system

for the automated translation of institutional academic material into ELF (but see Bernardini, 2014).

The translation memory resulting from the conversion of the CODE-UniBO-Par corpus into TMX format (cf. Section 4.3) may be used by non-native translators working with CAT tools as an aiding resource to translate Italian institutional academic (web-based) content into ELF. Future studies may evaluate the quality of language of the TM, e.g. by studying the terminology used in the bitexts and manually giving a quality mark to the various translation units. In the future, the resources created in this contribution may also be extended through the alignment of similar bitexts from different universities and academic years. These resources may eventually be used to develop an Italian-ELF web-based linguistic database for Italian academic institutions. Professional translators and university personnel may use it as a reference for the production of academic course descriptions in ELF. The database may also be integrated with existing or ad hoc glossaries in order to facilitate and speed up the production/translation process and guarantee terminology consistency at university and national (i.e. Italian) level.

Conclusion

The present study has been prompted by a practical need: the harmonization and internationalization process in contemporary higher education, which has led to a substantial increase in the demand for English translations in the institutional academic domain across Europe. Resources are consequently needed to help non-native higher education institutions translate cost-efficiently their websites into English.

In particular, the study has dealt with bitext sentence alignment and translation technology. In the first part of the study, a set of sentence aligners have been examined and evaluated according to a series of user-oriented parameters, i.e. software characteristics, basic features and advanced features. The analysis revealed that eAlign is the software that strikes the best balance between process automation and alignment accuracy. The tool is based on hunalign, which was deemed the most accurate sentence-level aligning algorithm available at the time of writing. It has been suggested that eAlign is particularly useful for the text pairs in the CODE-UniBO corpus (i.e. Italian-ELF academic course descriptions) and for items that display similar noisiness patterns (e.g. insertions, omissions, variation in structure, and so forth). The algorithm hunalign is also bundled with another aligner (i.e. InterText), which produces similar results to eAlign, but with a lower degree of process automation.

The second part of this study presented a heuristic approach to the creation and evaluation of Italian-ELF reference resources for non-native authors and translators working with English in the institutional academic domain. The approach adopted in this study consisted in the initial categorization of the CODE-UniBO corpus according to document length similarity, i.e. a set of uniform ranges representing the bitext variation in *characters*. The bitexts in each length variation range were then analyzed in terms of: (i) extent of parallelism, i.e. probability of translation equivalence between the *sentences* in the bitexts; (ii) segmentation and alignment accuracy, i.e. percentage of the number of correctly segmented and aligned *sentences* in the bitexts; and (iii) amount of leverageable content, i.e. percentage of correct *translation units* within a translation memory. The resulting data were analyzed at both sample and range level with the aim of exploring any existing connection between the results of these analyses and the bitext length

variation rates. Findings revealed that there exists a direct relation between the four aspects of the analysis, i.e. document length similarity, bitext parallelism, alignment accuracy, and resource leverageability. In fact, for each range the data obtained from the various analyses follow similar trends at sample level: they tend to increase as length variation rates decrease and approach 0, and to decrease as length variation rates increase, with few exceptions. On the other hand, the small difference among the data obtained for most ranges in the various analyses confirms the close relation between the aspects in (i), (ii), and (iii) also at range level. In particular, results revealed that bitexts containing longer Italian texts are likely to be more parallel than bitexts containing longer English texts. More specifically, results suggest that highly similar Italian-ELF bitexts (i.e. -15%/0% length variation rate) are more likely to yield the highest probability of translation equivalence in the sample. Accordingly, bitext containing longer Italian texts tend to display higher alignment accuracy and resource leverageability. However, in most cases alignment and leverageability results are still accurate enough in the case of longer English texts. This is particularly the case for bitexts containing several missing sections, which the aligner easily recognizes. Data obtained from the various analyses eventually led us to suggest that bitexts displaying 85% or higher parallelism rates usually return good results in terms of alignment accuracy and resource leverageability, i.e. greater than or equal to 85%.

As part of a larger project aimed at the creation of reference resources for non-native authors and translators in institutional academic settings (i.e. the CODE project), the last part of this study has presented the CODE-UniBO-Par corpus: a parallel corpus of 3,044 Italian-ELF web-based academic course descriptions that includes 1,269,072 Italian tokens (i.e. words) and 1,123,970 English tokens. The corpus was built through the automatic sentence alignment of all the bitexts in the CODE-UniBO corpus that satisfied the 85% or higher parallelism requirement. The CODE-UniBO-Par corpus was eventually exported to TMX format, resulting in 99,181 translation units.

Acknowledgments

This thesis would not have been possible without the suggestions, comments and corrections of Prof. Adriano Ferraresi and Prof. Silvia Bernardini, who have provided the author many inspiring ideas and helpful advices. Their contribution is sincerely appreciated and gratefully acknowledged. The author is also grateful to Eros Zanchetta for his technical help and support.

References

Afros, E., & Schryer, C. F. (2009). The genre of syllabus in higher education. *Journal of English for Academic Purposes, 8* (3), 224-233.

Altbach, P. G., & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education*, *11* (3-4), 290-305.

Berger, A. (2000). *The Align system.* Retrieved February 22, 2015, from http://web.archive.org/web/20031216214354/http://www-2.cs.cmu.edu/~aberger/software/align.html

Bernardini, S. (2014, June). *CODE: Cataloghi dell'Offerta Didattica in Europa. Sistematizzazione delle competenze, descrizione del genere e produzione in inglese lingua franca. Progetto FARB 2013-2015*. Poster presented at the meeting of the Department of Interpretation and Translation (DIT) of the University of Bologna, Bologna. Retrieved February 18, 2015 from http://code.sslmit.unibo.it/lib/exe/fetch.php?media=code_intro.pdf

Bernardini, S., Dalan, E., Ferraresi, A., Gaspari, F., Maldussi, D., Soffritti, M., Wiesmann, Zanchetta, E., & Zingaro, A., (2015). *CODE – Cataloghi dell'Offerta Didattica in Europa*. Retrieved February 05, 2015, from: http://code.sslmit.unibo.it/

Bernardini, S., Ferraresi, A., & Gaspari, F. (2010). Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language. In A. Linde López, & R. Crespo Jiménez (Eds.), *Professional English in the European context: The EHEA challenge* (pp. 27-53). Bern: Peter Lang.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Braune, F., & Fraser, A. (2010). Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. *Coling 2010: 23rd International Conference on Computational Linguistics. Posters Volume*, 81–89. Beijing.

Briel, D (n.d.). *bligner*. Retrieved February 22, 2015, from http://www.bligner.org/

BTS (2013). *Bologna Translation Service*. Retrieved February 04, 2015, from http://www.bologna-translation.eu/

Caiazzo, L. (2011). Hybridization in institutional language: Exploring *we* in the "About us" page of university websites. In S. Sarangi, V. Polese, & G. Caliendo (Eds.), *Genre(s) on the Move: Hybridization and Discourse Change in Specialized Communication* (pp. 243-260). Napoli: Edizioni Scientifiche Italiane.

Callahan, E. & Herring, S. C. (2012). Language choice on university websites: Longitudinal trends. *International Journal of Communication, 6*, 322-355.

Carey, R. (2014). *The ELFA project: Written academic ELF (WrELFA)*. Retrieved February 24, 2015, from http://www.helsinki.fi/englanti/elfa/wrelfa.html

Chuang, T. C., & Chang, J. S. (2002). Adaptive sentence alignment based on length and lexical information. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadephia, Pennsylvania, USA.

Dalan, E. (2012). *Costruzione automatica di corpora orientati al genere e fraseologia: Il caso delle guide web in inglese degli Atenei europei*. (Unpublished master's thesis). University of Bologna, Bologna.

Depraetere, H., Van den Bogaert, J., & Van de Walle, J. (2011). Bologna translation service: Online translation of course syllabi and study programmes in English. In M. L. Forcada, H. Depraetere, & V. Vandeghinste (Eds.), *Proceedings of the 15th conference of the European Association for Machine Translation* (pp. 29-34). Leuven, Belgium.

EHEA. (2014). *Bologna Process – European Higher Education Area*. Retrieved February 02, 2015, from http://www.ehea.info/

European Communities. (2009). *ECTS Users' Guide*. Retrieved February 18, 2015, from http://ec.europa.eu/education/tools/docs/ects-guide_en.pdf

Fairclough, N. (1993). Critical discourse analysis and the marketization of public discourse: the universities. *Discourse & Society, 4* (2), 133-168.

Farkas, A. (n.d.). *LF Aligner*. Retrieved February 22, 2015, from http://sourceforge.net/projects/aligner/

Ferraresi, A. & Bernardini, S. (2013). The academic web-as-corpus. In S. Evert, E. Stemle, & P. Rayson (Eds.), *Proceedings of the 8th Web as Corpus Workshop* (pp. 53-62). Lancaster, UK.

Ferraresi, A. & Bernardini, S. (forthcoming). Institutional academic English and its phraseology: Native and lingua franca perspectives.

Forcada, M. L., & Martin, R. (2010). *bitext2tmx: Bitext Aligner/Converter*. Retrieved February 22, 2015, from http://bitext2tmx.sourceforge.net/

GITS – Ginstrom IT Solutions (2008). *Align Assist Translation File Alignment Tool*. Retrieved February 22, 2015, from http://felix-cat.com/tools/align-assist/

Hofland, K., & Johansson, S. (1998). The Translation Corpus Aligner: A program for automatic alignment of parallel texts. In S. Johansson, & S. Oksefjell (Eds.), *Corpora and Crosslinguistic Research: Theory, Method, and Case Studies* (pp. 87-100). Amsterdam: Rodopi.

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

Jenkins, J. (2011). Accommodating (to) ELF in the international university. *Journal of Pragmatics, 43* (4), 926-936.

Kilgray. (2015). *memoQ 2014 R12 Help*. Retrieved February 22, 2015, from http://kilgray.com/memoq/2014R2/help-en/

Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.

Ma, X. (2006). Champollion: A Robust Parallel Text Sentence Aligner. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

Mautner, G. (2005). For-profit discourse in the nonprofit and public sectors. In G. Erreygers, & G. Jacobs (Eds.), *Language, communication and the economy* (pp. 25-44). Amsterdam: John Benjamins.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. In E. Brill, & K. Church (Eds.), *Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1–12). Philadelphia, Pennsylvania, USA.

MOKK Centre for Media Research and Education (2015). h*unalign – sentence aligner*. Retrieved February 22, 2015, from http://mokk.bme.hu/en/resources/hunalign/

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In S. D. Richardson (Ed.), *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA: Proceedings, volume 2499 of Lecture Notes in Computer Science*. Springer.

Morrish, L., & Sauntson, H. (2013). "Business-facing motors for economic development": An appraisal analysis of visions and values in the marketised UK university. *Critical Discourse Studies, 10* (1), 61-80.

Peter F. Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning sentences in parallel corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL)*, 169–176. Berkeley, California.

Sánchez Villamil, E., Sergio Ortiz-Rojas, Santos-Antón, Forcada, M. L., Simon, M., & Esplà, M. (2008). *Tagaligner: Aligner for parallel text*. Retrieved February 22, 2015, from http://tag-aligner.sourceforge.net/

Santos, D., & Oksefjell, S. (1999). Using a parallel corpus to validate independent claims. In H. Hasselgård, S. Johansson, & C. Fabricius-Hansen (Eds.), Information Structure in Parallel Texts. Special issue of *Languages in Contrast* 2:1 (pp. 115–130).

Seidlhofer, B. (2011). *Understanding English as a Lingua Franca: A Complete Introduction to the Theoretical Nature and Practical Implications of English used as a Lingua Franca*. Oxford: Oxford University Press.

Simard, M., Foster G., & Isabelle, P. (1992). Using cognates to align sentences in bilingual corpora. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, 67–81. Montreal, Canada.

Swales, J. M. (2004). *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.

Terminotix. (2008). *Alignment tools, AlignFactory*. Retrieved February 22, 2015, from http://www.terminotix.com/

Tiedemann, J. (2006). ISA & ICA - Two Web Interfaces for Interactive Alignment of Bitexts. *Proceedings of the 5th International Conference on Language Resources and Evaluation, (LREC 2006)*. Genoa, Italy.

Tiedemann, J. (2011). *Bitext alignment*. San Rafael, California: Morgan and Claypool Publishers.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages. *Proceedings of Recent Advances in Natural Language Processing (RANLP) 2005*, 590-596. Borovets, Bulgaria.

VOICE. (2013). *Vienna-Oxford International Corpus of English*. Retrieved February 04, 2015, from https://www.univie.ac.at/voice/

Vondřička, P. (2014a). Aligning parallel texts with InterText. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 1875-1879. European Language Resources Association (ELRA).

Vondřička, P. (2014b). *InterText*. Retrieved February 22, 2015, from http://wanthalf.saga.cz/intertext

William A. Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, 177-184. Berkeley, California.

Wu, D. (1994). Aligning a parallel English-Chinese corpus statistically with lexical criteria. *Proceedings of the 32rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 80–87. New Mexico State University.

Wu, D. (2010). Alignment. In N. Indurkhya, & F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, second edition (pp. 367–408). CRC Press.

Appendix A | Bitext sentence aligners

In the following table, the set of aligners examined in Chapter 2 will be compared to each other according to a series of user-oriented parameters, i.e. software characteristics, basic features and advanced features (see Section 2.1). The aligners are divided in three categories, reported in brackets in the column ALIGNER on the left-hand side of the table: (1) aligning algorithms and CLI aligners; (2) free GUI alignment tools; and (3) commercial GUI alignment tools.

ALIGNER (1)	SOFTWARE CHARACTERISTICS				BASIC FEATURES					ADVANCED FEATURES		
	APPROACH	PROGRAMMING LANGUAGE AND REQUIREMENTS	INTERFACE	AVAILABILITY	GRANULARITY LEVEL	LANGUAGE PAIRS	FILE PAIRS PER ALIGNMENT	BITEXT LINK CORRESPONDENCE	OUTPUT FORMAT(S)	ADDITIONAL RESOURCES	PRE-PROCESSING	SEGMENTATION RULES CONFIGURATION
Gale & Church aligning algorithm	Length-based (characters)	Perl	CLI	Freely available at http://goo.gl/D5LcqX (last visited February 23, 2015)	Sentence	Indo-European languages	n/a	Strong preference for 1-1 correspondences. Other correspondences include 0-1, 1-1, 2-1, 1-2, 2-2	n/a	Not required	Corpus pre-segmentation at paragraph level (*recommended*)	No
Moore's bilingual sentence aligner	Length-based (characters) + lexical information (words)	Perl	CLI	Freely available at http://goo.gl/qonnjA (last visited February 23, 2015)	Sentence	Language independent	n/a	1-1	n/a	Not required	Corpus pre-segmentation at sentence and word level: one sentence per line, space-delimited words	No
CTK: Champollion Tool Kit	Lexical information (words) + length information (characters)	Perl	CLI	Freely available at http://goo.gl/xUIvfP (last visited February 23, 2015)	Sentence	EN-AR, EN-ZH, EN-HI	n/a	The aligner assumes noisy input: 1-0, 0-1, 1-1, *n*-1, 1-*n*	Aligned sentence blocks, one per line: `language1 sentence ids <=> language2 sentence ids`	**Bilingual lexicon**	Sentence-segmented input files: one sentence per line. Word segmentation improves both precision and recall	No
TCA: Translation Corpus Aligner	Lexical-information (anchor words) + length information (characters) + cognates, punctuation, structural and formatting information	Java Texts have to be marked up in XML	n/a Interactive aligner	Not currently available. Product description: http://goo.gl/vPNpM4 (last visited February 23, 2015)	Sentence	Language dependent. Tested on EN-NO, EN-DE, EN-NL, EN-ES and EN-PT	n/a	1-0, 1-1, 1-2 1-0 and 1-2 have to be checked manually	Several output formats, TEI and ParaConc / Multiconcord text	**Bilingual lexicon** (referred to as *anchor list*)	Sentence-segmented input files	n/a
GMA	Geometric Mapping (SIMR) and Alignment (GSA). Lexical information (words) + length	Java Only tested on Linux/i386 and Solaris/SPARC)	CLI	Freely available at http://goo.gl/fJdsAl (last visited February 23, 2015)	Paragraph, sentence	Tested on FR-EN, ES-EN, KO-EN, ZH-EN, AR-EN, CS-EN, MS-EN, RU-EN	n/a	1-0, 0-1, 1-1, 2-1, 1-2, 2-2	Aligned sentence blocks, one per line	**Bilingual lexicon**; lists of **stop words** (*recommended*)	Not required	No

ALIGNER (1)	SOFTWARE CHARACTERISTICS				BASIC FEATURES					ADVANCED FEATURES		
	APPROACH	PROGRAMMING LANGUAGE AND REQUIREMENTS	INTERFACE	AVAILABILITY	GRANULARITY LEVEL	LANGUAGE PAIRS	FILE PAIRS PER ALIGNMENT	BITEXT LINK CORRESPONDENCE	OUTPUT FORMAT(S)	ADDITIONAL RESOURCES	PRE-PROCESSING	SEGMENTATION RULES CONFIGURATION
hunalign	Length-based (characters) + lexical information (words)	C++	CLI	Freely available at http://goo.gl/hCStqv (last visited February 23, 2015)	Sentence	Language independent	n/a	*n*-0, 0-*n*, 1-1, *n*-1, 1-*n*	**Text** or **ladder format**: tab-separated or newline-separates sentence links + confidence value	**Dictionary** (*recommended*): newline-separated items. **partialAlign** required with 10K+ sentences	Tokenized and sentence-segmented input files: one sentence per line, space-delimited words	No
Tag-aligner	XML-based tag structure and length information	Requirements: - i686, ppc, SPARC, etc. - g++ (version 2.95 or newer) - GNU make - XML-like markup-language-based input files	CLI	Freely available at http://goo.gl/Jc7VEJ (last visited February 23, 2015)	Sentence	n/a	n/a	n/a	TMX	Not required	n/a	n/a
Align	Dynamic programming + lexical information (anchor words)	C++ Meant to be compiled within a Unix environment	CLI	Freely available at http://goo.gl/hrvQwg (last visited February 23, 2015)	Paragraph, sentence, sub-sentence	n/a Initially developed for FR-EN	n/a	1-0, 0-1, 1-1, 2-1, 1-2	n/a	The **scoring function** and other **code functions** have to be modified	Tokenized and sentence-segmented input files: one sentence per line, space-delimited words	No
Gargan-tua	n/a	Perl	CLI	Freely available at http://goo.gl/eNx24x (last visited February 23, 2015)	Sentence	Language independent	n/a	n/a The aligner is intended for symmetrical and asymmetrical parallel corpora	n/a	Not required	Tokenized and sentence-segmented input files: one sentence per line, space-delimited words.	No
bligner	n/a	Perl and Python	CLI	Freely available at http://goo.gl/OHxPcp (last visited February 23, 2015)	Paragraph, sentence	n/a	n/a	n/a	TMX	Not required	Sentence-segmented input files	Yes

ALIGNER (2)	SOFTWARE CHARACTERISTICS				BASIC FEATURES					ADVANCED FEATURES		
	APPROACH	PROGRAMMING LANGUAGE AND REQUIREMENTS	INTERFACE	AVAILABILITY	GRANULARITY LEVEL	LANGUAGE PAIRS	FILE PAIRS PER ALIGNMENT	BITEXT LINK CORRESPONDENCE	OUTPUT FORMAT(S)	ADDITIONAL RESOURCES	PRE-PROCESSING	SEGMENTATION RULES CONFIGURATION
LF aligner	Based on **hunalign**: length-based (characters) + lexical information (words)	Perl. The Windows version is packaged into a standalone EXE file	GUI (on Windows) Interactive editor	Freely available at http://goo.gl/JkynGX (last visited February 23, 2015)	Sentence	180+ languages	1	*n*-0, 0-*n*, 1-1, *n*-1, 1-*n*	TMX 1.4, XLS, tab-delimited TXT	Not required. **Built-in dictionary data** for 800+ combinations of 32 languages	Not required	No
InterText Editor	Based on **hunalign**: length-based (characters) + lexical information (words)	C++	GUI Interactive editor	Freely available at http://goo.gl/Nu0Pcr (last visited February 23, 2015)	Paragraph, sentence	Language independent	1 It handles multiple file pairs	*n*-0, 0-*n*, 1-1, *n*-1, 1-*n*	TMX 1.4b, ParaConc text, pre-defined profiles for newline aligned texts	Not required	Not required Recommended: sentence-segmented input files, one sentence per line	Integrated, fully configurable sentence splitter. based on regular expressions
eAlign (formerly known as **Super-Align**)	Based on **hunalign**: length-based (characters) + lexical information (words)	n/a	GUI Interactive editor	Freely available at http://goo.gl/avUer4 (last visited February 23, 2015)	Paragraph, sentence	*Generic* option + 19 specific languages: CS, DA, NL, EN, FR, DE, EL, GU, HE, HU, IT, MX, NO, PL, PO, RU, ES, SV, ZH	1 It handles multiple file pairs	*n*-0, 0-*n*, 1-1, *n*-1, 1-*n*	TMX 1.4, CSV, tab-delimited TXT, DOC in tables Output files can be merged into a single TMX file	Not required	Not required	Fully configurable segmentation rules for each language
Align Assist (Felix)	n/a	n/a	GUI (part of the CAT tool Felix) Interactive editor	Freely available at http://goo.gl/jcHetu (last visited February 23, 2015)	Sentence	50+ languages	1	1-1	Felix TM format, TMX 1.4	Not required	Not required	No
bitext-2tmx	n/a	Java	GUI	Freely available at http://goo.gl/cze7Wz (last visited February 23, 2015)	Sentence	21 languages: AR, CA, ZH, CS, DA, NL, EN, FI, FR, DE, HU, IT, JA, KO, NO, PL, PT, RU, ES, SV, TH	1	1-1, 2-1, 1-2 Low output quality	TMX 1.1	Not required	Not required Recommended: sentence-segmented input files	No

ALIGNER (3)	SOFTWARE CHARACTERISTICS				BASIC FEATURES					ADVANCED FEATURES		
	APPROACH	PROGRAMMING LANGUAGE AND REQUIREMENTS	INTERFACE	AVAILABILITY	GRANULARITY LEVEL	LANGUAGE PAIRS	FILE PAIRS PER ALIGNMENT	BITEXT LINK CORRESPONDENCE	OUTPUT FORMAT(S)	ADDITIONAL RESOURCES	PRE-PROCESSING	SEGMENTATION RULES CONFIGURATION
AlignFactory Light	n/a	n/a	GUI Interactive editor	Commercial product: http://goo.gl/2GExbe (last visited February 23, 2015)	Paragraph, sentence	Language independent	More than 1 Automatic pairing of files not available in the **Light** version	1-1, 2-1, 1-2 1-0 and 0-1 are automatically merged with other segments	TMX, HTML and XML LogiTerm Bitext	Not required	Not required	Yes
SDL Trados Studio 2014	n/a	n/a	GUI The aligner is part of the CAT tool	Commercial product http://goo.gl/2lVp8Z (last visited February 23, 2015)	Sentence	Language independent	More than 1 Automatic pairing of files	1-1, 2-1, 1-2	TMX Output files are merged into a single file	Not required	Not required	Yes
memoQ 2014 R2	LiveAlign technology: statistical + linguistic algorithms, structural information, anchor terms, formatting information, inline tags	n/a	GUI The aligner is part of the CAT tool	Commercial product http://goo.gl/zzaJBb (last visited February 23, 2015)	Sentence	Language independent	More than 1 Automatic pairing of files	1-1, 2-1, 1-2, 2-2 1-0 and 0-1 can be marked manually and blocked for automatic re-alignment	TMX 1.4 Output files are merged into a single file	Not required Users can add **term bases** to use entries as anchors	Not required	Yes (regular expressions)

Appendix B | Document length similarity and bitext parallelism in the sample

LENGTH VARIATION RANGE: -35% ≤ δ > -25%								
NUM_CHAR1 [EN]	NUM_CHAR 2 [IT]	DIFF_CHAR	LENGTH VARIATION RATE	NUM_SENT1 [EN]	NUM_SENT2 [IT]	TOT_SENT	TOT_SENT_PARAL-LEL	BITEXT PARALLELISM
1094	1945	-851	**-28.00263244**	18	26	44	31	**0.704545455**
1331	2385	-1054	**-28.36383208**	35	39	74	64	**0.864864865**
1065	1822	-757	**-26.22099065**	16	19	35	32	**0.914285714**
989	1800	-811	**-29.07852277**	29	39	68	59	**0.867647059**
487	1008	-521	**-34.84949833**	12	16	28	24	**0.857142857**
2225	4168	-1943	**-30.39261692**	26	30	56	49	**0.875**
909	1586	-677	**-27.13426854**	23	29	52	44	**0.846153846**
1418	2804	-1386	**-32.82804358**	21	32	53	42	**0.79245283**
1038	1806	-768	**-27.00421941**	17	20	37	34	**0.918918919**
1453	2507	-1054	**-26.61616162**	36	42	78	64	**0.820512821**
737	1288	-551	**-27.20987654**	16	19	35	32	**0.914285714**
672	1191	-519	**-27.85829308**	21	21	42	36	**0.857142857**
407	859	-452	**-35.70300158**	14	18	32	27	**0.84375**
1417	2448	-1031	**-26.67529107**	36	39	75	64	**0.853333333**
2834	5274	-2440	**-30.09373458**	35	41	76	63	**0.828947368**
		AVG	**-29.20206555**				**AVG**	**0.850598909**

LENGTH VARIATION RANGE: -25% ≤ δ > -15%								
NUM_CHAR1 [EN]	NUM_CHAR 2 [IT]	DIFF_CHAR	LENGTH VARIATION RATE	NUM_SENT1 [EN]	NUM_SENT2 [IT]	TOT_SENT	TOT_SENT_PARAL-LEL	BITEXT PARALLELISM
333	515	-182	**-21.46226415**	12	12	24	22	**0.916666667**
2635	4451	-1816	**-25.62799887**	27	38	65	55	**0.846153846**
1450	2026	-576	**-16.570771**	26	29	55	0	**0**
3438	5049	-1611	**-18.98197243**	24	29	53	40	**0.754716981**
1171	1694	-523	**-18.2547993**	39	43	82	76	**0.926829268**
2318	3341	-1023	**-18.07739883**	27	31	58	55	**0.948275862**
1116	1895	-779	**-25.87180339**	26	31	57	51	**0.894736842**
2203	3627	-1424	**-24.42538593**	29	34	63	57	**0.904761905**
1677	2634	-957	**-22.19902575**	28	32	60	58	**0.966666667**
908	1442	-534	**-22.72340426**	16	17	33	31	**0.939393939**
1616	2365	-749	**-18.81436825**	19	19	38	37	**0.973684211**
3064	4610	-1546	**-20.14594735**	55	55	110	96	**0.872727273**
3651	5078	-1427	**-16.34780616**	47	44	91	85	**0.934065934**
1696	2386	-690	**-16.90347869**	29	33	62	54	**0.870967742**
1750	2755	-1005	**-22.30854606**	23	25	48	41	**0.854166667**
		AVG(0s)	**-20.58099803**				**AVG(0s)**	**0.840254254**
		AVG	**-20.86744282**				**AVG**	**0.900272415**

LENGTH VARIATION RANGE: -15% ≤ δ > -5%								
NUM_CHAR1 [EN]	NUM_CHAR 2 [IT]	DIFF_CHAR	LENGTH VARIATION RATE	NUM_SENT1 [EN]	NUM_SENT2 [IT]	TOT_SENT	TOT_SENT_PARAL-LEL	BITEXT PARALLELISM
1356	1831	-475	**-14.90429871**	16	18	34	32	**0.941176471**
2466	3024	-558	**-10.16393443**	49	46	95	94	**0.989473684**
2498	3205	-707	**-12.39698404**	29	31	60	54	**0.9**
1399	1679	-280	**-9.096816114**	32	33	65	63	**0.969230769**
3175	4061	-886	**-12.24433389**	78	78	156	156	**1**
1766	2420	-654	**-15.62350693**	21	25	46	44	**0.956521739**
2232	2524	-292	**-6.13961312**	48	47	95	94	**0.989473684**
1944	2454	-510	**-11.59618008**	40	42	82	75	**0.914634146**
2811	3384	-573	**-9.249394673**	43	40	83	80	**0.963855422**
5851	7615	-1764	**-13.0996584**	64	64	128	111	**0.8671875**
2029	2612	-583	**-12.56194786**	74	85	159	144	**0.905660377**
10090	11805	-1715	**-7.832838548**	205	208	413	407	**0.985472155**
1445	1973	-528	**-15.44763019**	20	24	44	42	**0.954545455**
2029	2314	-285	**-6.562284135**	30	31	61	60	**0.983606557**
2352	2774	-422	**-8.232539992**	26	26	52	52	**1**
		AVG	**-11.01013074**				**AVG**	**0.954722531**

LENGTH VARIATION RANGE: -5% ≤ δ ≥ +5%								
NUM_CHAR1 [EN]	NUM_CHAR 2 [IT]	DIFF_CHAR	LENGTH VARIATION RATE	NUM_SENT1 [EN]	NUM_SENT2 [IT]	TOT_SENT	TOT_SENT_PARAL-LEL	BITEXT PARALLELISM
4822	5081	-259	**-2.61536908**	45	45	90	85	**0.944444444**
6935	7059	-124	**-0.88609404**	44	44	88	88	**1**
1438	1514	-76	**-2.574525745**	30	27	57	52	**0.912280702**
2078	2271	-193	**-4.437801794**	60	60	120	120	**1**
4932	4904	28	**0.284668564**	39	40	79	0	**0**
1763	1589	174	**5.190930788**	22	21	43	38	**0.88372093**
14182	15467	-1285	**-4.33404162**	136	139	275	268	**0.974545455**
11857	11848	9	**0.037966674**	57	57	114	114	**1**
1626	1815	-189	**-5.492589364**	22	21	43	42	**0.976744186**
6615	7457	-842	**-5.98351336**	92	91	183	182	**0.994535519**
3518	3317	201	**2.940746159**	75	74	149	0	**0**
1492	1459	33	**1.118264995**	30	29	59	56	**0.949152542**
2139	2174	-35	**-0.811500116**	21	20	41	41	**1**
2570	2569	1	**0.019459039**	31	31	62	0	**0**
2444	2279	165	**3.49354224**	26	23	49	44	**0.897959184**
		AVG(0s)	**-0.936657111**				**AVG(0s)**	**0.768892197**
		AVG	**-1.168811694**				**AVG**	**0.957580269**

LENGTH VARIATION RANGE: +5% < δ ≥ +15%								
NUM_CHAR1 [EN]	NUM_CHAR 2 [IT]	DIFF_CHAR	LENGTH VARIATION RATE	NUM_SENT1 [EN]	NUM_SENT2 [IT]	TOT_SENT	TOT_SENT_PARAL-LEL	BITEXT PARALLELISM
2232	1735	497	**12.52835896**	35	29	64	0	**0**
5987	5046	941	**8.528958579**	71	61	132	114	**0.863636364**
2135	1685	450	**11.78010471**	39	36	75	71	**0.946666667**
2463	2000	463	**10.37418777**	54	51	105	96	**0.914285714**
2478	2167	311	**6.695371367**	33	27	60	53	**0.883333333**
4173	3504	669	**8.71434154**	71	68	139	0	**0**
2726	2368	358	**7.027875932**	50	48	98	0	**0**
777	569	208	**15.45319465**	16	15	31	30	**0.967741935**
7020	5603	1417	**11.22554068**	78	71	149	139	**0.932885906**
2632	2213	419	**8.648090815**	57	55	112	85	**0.758928571**
1566	1314	252	**8.75**	34	32	66	64	**0.96969697**
1598	1393	205	**6.853895018**	18	17	35	0	**0**
3175	2682	493	**8.41727847**	52	38	90	76	**0.844444444**
4588	3638	950	**11.54874787**	42	36	78	0	**0**
808	675	133	**8.968307485**	20	19	39	0	**0**
		AVG(0s)	**9.700950257**				**AVG(0s)**	**0.53877466**
		AVG	**9.98585856**				**AVG**	**0.897957767**

LENGTH VARIATION RANGE: +15% < δ ≥ +25%								
NUM_CHAR1 [EN]	NUM_CHAR 2 [IT]	DIFF_CHAR	LENGTH VARIATION RATE	NUM_SENT1 [EN]	NUM_SENT2 [IT]	TOT_SENT	TOT_SENT_PARAL-LEL	BITEXT PARALLELISM
1518	1097	421	**16.09942639**	26	24	50	0	**0**
1424	1014	410	**16.81706317**	15	14	29	27	**0.931034483**
3264	2351	913	**16.26001781**	59	52	111	0	**0**
2568	1804	764	**17.47483989**	31	40	71	57	**0.802816901**
4060	2884	1176	**16.93548387**	54	41	95	0	**0**
2865	2026	839	**17.15395625**	39	37	76	0	**0**
3528	2420	1108	**18.62811029**	63	47	110	92	**0.836363636**
2378	1655	723	**17.92710141**	32	27	59	0	**0**
2490	1774	716	**16.79174484**	33	26	59	0	**0**
917	565	352	**23.75168691**	22	19	41	38	**0.926829268**
700	451	249	**21.63336229**	20	18	38	36	**0.947368421**
2190	1557	633	**16.89351481**	35	28	63	0	**0**
918	654	264	**16.79389313**	18	16	34	32	**0.941176471**
2212	1365	847	**23.67906067**	49	23	72	0	**0**
965	650	315	**19.50464396**	21	19	40	36	**0.9**
		AVG(0s)	**18.42292705**				**AVG(0s)**	**0.419039279**
		AVG	**19.22908566**				**AVG**	**0.897941311**

LENGTH VARIATION RANGE: +25% < δ ≥ +35%								
NUM_ CHAR1 [EN]	NUM_ CHAR 2 [IT]	DIFF_ CHAR	LENGTH VARIATION RATE	NUM_ SENT1 [EN]	NUM_ SENT2 [IT]	TOT_ SENT	TOT_SENT _PARAL- LEL	BITEXT PARALLELISM
1573	878	695	**28.35577315**	27	19	46	36	**0.782608696**
2367	1285	1082	**29.62760131**	27	23	50	26	**0.52**
1094	526	568	**35.0617284**	24	19	43	38	**0.88372093**
2191	1089	1102	**33.59756098**	38	29	67	0	**0**
2834	1470	1364	**31.69144981**	33	25	58	47	**0.810344828**
729	420	309	**26.89295039**	17	13	30	26	**0.866666667**
2698	1457	1241	**29.86762936**	29	18	47	25	**0.531914894**
1614	799	815	**33.77538334**	26	21	47	0	**0**
1048	604	444	**26.87651332**	19	17	36	0	**0**
3650	1828	1822	**33.26031398**	40	20	60	39	**0.65**
		AVG(0s)	**30.9006904**				**AVG(0s)**	**0.504525601**
		AVG	**30.6796352**				**AVG**	**0.720750859**

Appendix C | Bitext automatic segmentation and alignment[26]

LENGTH VARIATION RANGE: -35% ≤ δ > -25%					
TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT _ALIGN	ALIGNMENT ACCURACY	NOTE
44	27	**0.613636364**	33	**0.75**	
74	68	**0.918918919**	69	**0.932432432**	
35	35	**1**	35	**1**	
68	39	**0.573529412**	31	**0.455882353**	*Missing punctuation*
28	24	**0.857142857**	28	**1**	
56	50	**0.892857143**	54	**0.964285714**	
52	50	**0.961538462**	50	**0.961538462**	
53	49	**0.924528302**	48	**0.905660377**	
37	37	**1**	37	**1**	
78	70	**0.897435897**	70	**0.897435897**	
35	33	**0.942857143**	35	**1**	
42	36	**0.857142857**	35	**0.833333333**	
32	28	**0.875**	32	**1**	
75	39	**0.52**	68	**0.906666667**	*Missing punctuation*
76	68	**0.894736842**	60	**0.789473684**	
	AVG	**0.848621613**	**AVG**	**0.893113928**	

LENGTH VARIATION RANGE: -25% ≤ δ > -15%					
TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT _ALIGN	ALIGNMENT ACCURACY	NOTE
24	24	**1**	22	**0.916666667**	
65	45	**0.692307692**	54	**0.830769231**	*Bibliographic ref.*
					English bitext
53	47	**0.886792453**	41	**0.773584906**	
82	78	**0.951219512**	71	**0.865853659**	
58	56	**0.965517241**	56	**0.965517241**	
57	53	**0.929824561**	46	**0.807017544**	
63	51	**0.80952381**	55	**0.873015873**	
60	41	**0.683333333**	47	**0.783333333**	*Missing punctuation*
33	33	**1**	33	**1**	
38	32	**0.842105263**	38	**1**	
110	106	**0.963636364**	93	**0.845454545**	
91	74	**0.813186813**	83	**0.912087912**	*Bibliographic ref.*
62	59	**0.951612903**	57	**0.919354839**	
48	37	**0.770833333**	48	**1**	
	AVG	**0.875706663**	**AVG**	**0.892332554**	

[26] Grey-shaded empty rows correspond to pairs of English texts or automatically translated texts (see Section 4.1.1).

LENGTH VARIATION RANGE: -15% ≤ δ > -5%					
TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT _ALIGN	ALIGNMENT ACCURACY	NOTE
34	34	**1**	34	**1**	
95	92	**0.968421053**	94	**0.989473684**	
60	42	**0.7**	54	**0.9**	*Missing punctuation*
65	26	**0.4**	58	**0.892307692**	*Missing punctuation*
156	38	**0.243589744**	156	**1**	*Missing punctuation*
46	43	**0.934782609**	44	**0.956521739**	
95	95	**1**	95	**1**	
82	60	**0.731707317**	50	**0.609756098**	
83	81	**0.975903614**	80	**0.963855422**	*Bibliographic ref.*
128	109	**0.8515625**	94	**0.734375**	
159	150	**0.943396226**	129	**0.811320755**	
413	383	**0.927360775**	402	**0.973365617**	
44	44	**1**	44	**1**	
61	42	**0.68852459**	58	**0.950819672**	*Missing punctuation*
52	52	**1**	52	**1**	
	AVG	**0.824349895**	**AVG**	**0.918786379**	

LENGTH VARIATION RANGE: -5% ≤ δ ≥ +5%					
TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT _ALIGN	ALIGNMENT ACCURACY	NOTE
90	75	**0.833333333**	75	**0.833333333**	
88	79	**0.897727273**	86	**0.977272727**	
57	52	**0.912280702**	54	**0.947368421**	
					Machine translation
					English bitext
43	43	**1**	39	**0.906976744**	
275	131	**0.476363636**	265	**0.963636364**	*Bibliographic ref.*
114	108	**0.947368421**	110	**0.964912281**	
43	41	**0.953488372**	42	**0.976744186**	
183	178	**0.972677596**	181	**0.989071038**	
					English bitext
59	31	**0.525423729**	57	**0.966101695**	*Missing punctuation*
41	41	**1**	41	**1**	
					English bitext
49	47	**0.959183673**	42	**0.857142857**	
	AVG	**0.86162243**	**AVG**	**0.943869059**	

LENGTH VARIATION RANGE: +5% < δ ≥ +15%					
TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT _ALIGN	ALIGNMENT ACCURACY	NOTE
					English bitext
132	114	**0.863636364**	109	**0.825757576**	
75	75	**1**	71	**0.946666667**	
105	52	**0.495238095**	98	**0.933333333**	*Missing punctuation*
60	55	**0.916666667**	56	**0.933333333**	
					English bitext
					English bitext
31	31	**1**	31	**1**	
149	133	**0.89261745**	145	**0.973154362**	
112	102	**0.910714286**	70	**0.625**	
66	66	**1**	64	**0.96969697**	
					English bitext
90	80	**0.888888889**	85	**0.944444444**	
					English bitext
					English bitext
	AVG	**0.885306861**	**AVG**	**0.905709632**	

LENGTH VARIATION RANGE: +15% < δ ≥ +25%					
TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT _ALIGN	ALIGNMENT ACCURACY	NOTE
					English bitext
29	27	**0.931034483**	28	**0.965517241**	
					English bitext
71	56	**0.788732394**	53	**0.746478873**	
					English bitext
					English bitext
110	54	**0.490909091**	104	**0.945454545**	*Bibliographic ref.*
					English bitext
					English bitext
41	41	**1**	41	**1**	
38	38	**1**	38	**1**	
					English bitext
34	28	**0.823529412**	34	**1**	*Missing punctuation*
					English bitext
40	40	**1**	38	**0.95**	
	AVG	**0.86202934**	**AVG**	**0.943921523**	

LENGTH VARIATION RANGE: +25% < δ ≥ +35%					
TOT_SENT	TOT_SENT _SEGM	SEGMENTATION ACCURACY	TOT_SENT _ALIGN	ALIGNMENT ACCURACY	NOTE
46	46	**1**	44	**0.956521739**	
50	36	**0.72**	25	**0.5**	
43	39	**0.906976744**	43	**1**	
					English bitext
58	58	**1**	56	**0.965517241**	
30	26	**0.866666667**	30	**1**	
47	33	**0.70212766**	27	**0.574468085**	
					English bitext
					English bitext
60	28	**0.466666667**	53	**0.883333333**	*Missing punctuation*
	AVG	**0.808919677**	**AVG**	**0.8399772**	

Appendix D | Resource leverageability(i.e. TM accuracy)[27]

LENGTH VARIATION RANGE: -35% ≤ δ > -25%			
TOT_TU	TOT_TU_CORRECT	TM ACCURACY	NOTE
12	9	**0.75**	
28	27	**0.964285714**	
13	13	**1**	
20	10	**0.5**	*Missing punctuation*
9	9	**1**	
20	19	**0.95**	
20	19	**0.95**	
18	17	**0.944444444**	
14	14	**1**	
31	29	**0.935483871**	
13	13	**1**	
16	14	**0.875**	
8	8	**1**	
17	14	**0.823529412**	*Missing punctuation*
30	24	**0.8**	
	AVG	**0.899516229**	

LENGTH VARIATION RANGE: -25% ≤ δ > -15%			
TOT_TU	TOT_TU_CORRECT	TM ACCURACY	NOTE
9	8	**0.888888889**	
26	22	**0.846153846**	*Bibliographic references*
			English bitext
18	13	**0.722222222**	
36	32	**0.888888889**	
24	23	**0.958333333**	
23	19	**0.826086957**	
22	21	**0.954545455**	
16	15	**0.9375**	*Missing punctuation*
12	12	**1**	
12	12	**1**	
44	38	**0.863636364**	
41	38	**0.926829268**	*Bibliographic references*
25	23	**0.92**	
14	14	**1**	
	AVG	**0.909506087**	

[27] Grey-shaded empty rows correspond to pairs of English texts or automatically translated texts (see Section 4.1.1).

LENGTH VARIATION RANGE: -15% ≤ δ > -5%			
TOT_TU	**TOT_TU_CORRECT**	**TM ACCURACY**	**NOTE**
13	13	**1**	
42	42	**1**	
19	17	**0.894736842**	*Missing punctuation*
13	12	**0.923076923**	*Missing punctuation*
21	21	**1**	*Missing punctuation*
18	18	**1**	
43	43	**1**	
31	20	**0.64516129**	
35	34	**0.971428571**	*Bibliographic references*
51	39	**0.764705882**	
69	58	**0.84057971**	
194	191	**0.984536082**	
17	17	**1**	
21	21	**1**	*Missing punctuation*
23	23	**1**	
	AVG	**0.934948353**	

LENGTH VARIATION RANGE: -5% ≤ δ ≥ +5%			
TOT_TU	**TOT_TU_CORRECT**	**TM ACCURACY**	**NOTE**
38	33	**0.868421053**	
37	37	**1**	
23	23	**1**	
			Machine Translation
			English bitext
18	16	**0.888888889**	
209	208	**0.995215311**	*Bibliographic references*
52	52	**1**	
18	18	**1**	
81	81	**1**	
			English bitext
14	13	**0.928571429**	*Missing punctuation*
17	17	**1**	
			English bitext
20	17	**0.85**	
	AVG	**0.957372426**	

LENGTH VARIATION RANGE: +5% < δ ≥ +15%			
TOT_TU	**TOT_TU_CORRECT**	**TM ACCURACY**	**NOTE**
			English bitext
50	45	**0.9**	
32	31	**0.96875**	
26	24	**0.923076923**	*Missing punctuation*
23	22	**0.956521739**	
			English bitext
			English bitext
12	12	**1**	
64	62	**0.96875**	
40	24	**0.6**	
28	28	**1**	
			English bitext
33	32	**0.96969697**	
	AVG	**0.92075507**	

LENGTH VARIATION RANGE: +15% < δ ≥ +25%			
TOT_TU	**TOT_TU_CORRECT**	**TM ACCURACY**	**NOTE**
			English bitext
10	10	**1**	
			English bitext
27	21	**0.777777778**	
			English bitext
			English bitext
26	26	**1**	*Bibliographic references*
			English bitext
			English bitext
16	16	**1**	
15	15	**1**	
			English bitext
11	11	**1**	*Missing punctuation*
16	15	**0.9375**	
	AVG	**0.959325397**	

LENGTH VARIATION RANGE: +25% < δ ≥ +35%			
TOT_TU	**TOT_TU_CORRECT**	**TM ACCURACY**	**NOTE**
16	15	**0.9375**	
14	7	**0.5**	
16	16	**1**	
			English bitext
20	19	**0.95**	
10	10	**1**	
12	7	**0.583333333**	
			English bitext
			English bitext
11	10	**0.909090909**	*Missing punctuation*
	AVG	**0.839989177**	