Understanding the composition of parallel corpora from the web

Marco Brunello
Centre for Translation Studies
University of Leeds
Leeds, LS2 9JT, United Kingdom
mlmb@leeds.ac.uk

ABSTRACT
Although it is fundamental to have a good fit between the text typology of training data and to-be-translated data in machine translation, there is a lack of studies on analysing parallel data under this point of view. This paper describes some studies made with the aim of understanding the composition of parallel corpora, in particular by using topic modeling.

Keywords
web as corpus, machine translation, parallel corpora, topic modeling

1. INTRODUCTION
Parallel texts have always been employed in translation studies, and in the last few decades large progresses in computational linguistics led the sub-field of machine translation to make wide use of large collection of aligned parallel corpora. Particularly those paradigms that strongly rely on the exploitation of this kind of resources to be used as training data, such as statistical machine translation (SMT), took advantage of this possibility. But, if the building of monolingual resources from the web comes with some undeniable issues - unknown overall composition of the Internet, continuous contents changing, unbalancedness towards particularly widespread languages or text varieties - the operation of retrieving multilingual data requires further technical skills for locating and pairing translated texts. As a consequence, the research in this field has not prospered in the same measure of monolingual web-as-corpus linguistics. Nevertheless, the use of the web as a source for parallel corpora is not an inert field of study. There is a quite rich tradition of studies in this direction, which includes a number of automatic tools that have been developed by several research groups, able to retrieve parallel data from the web (see Section 3.1).

However, it seems that not enough attention has been put on understanding what kind of documents are contained in parallel corpora in relation to their text variety. On one hand it is clear that, in a situation of scarceness of publicly available parallel corpora (especially when dealing with systems that base their successfulness on the employment of large collections of bitexts), the rule "more data is better data" is applied. On the other hand it could be still useful to have a clear knowledge and control of the kind of text to be translated using parallel data, especially in relation to their text variety, whatever is the chosen technology.

Even just the use of topic modeling techniques [14], employing an unsupervised method able to understand the composition of a corpus guessing what are the topic domains contained in a corpus, could be enough to give an idea of what are the kind of documents particularly interesting or useful for a particular task, such as the creation of translation memories to be used for fuzzy matches in a computer-assisted translation program or the selection of the most suitable training data for SMT. In this paper some exploratory studies conducted on two English-Italian parallel corpora are presented. In the first case, an already existing and very a well known resource, Europarl [7], has been used, while in the second case the employment of an automatic collection of translated texts from the web is described in order to collect data that will be part of a new parallel corpus. In this second case, the analysis will also help the reader to understand which topics are covered by the parallel web, and their potential usefulness in translation tasks.

The remainder of this document is organized as follows: Section 2 shows the experiments done on Europarl; Section 3, after showing previous related works on the collection of parallel corpora from the web, presents the way in which some of these techniques have been applied to collect part of a general-purpose English-Italian parallel web corpus, and the related topic modeling analysis; in Section 4 future directions of the results are suggested.

2. EXPERIMENTS ON EUROPARL
2.1 Topic modeling
As mentioned before, when talking about parallel corpora, Europarl is a very well known resource, widely employed by SMT scholars and researchers, since it provides aligned parallel texts up to 50 million words per language for the main European languages. Although it has been used in a variety of experiments, to our knowledge none of them has tried to explore its composition in terms of the arguments discussed in it. Being its parallel texts transcriptions of proceedings of

the European parliament, the supposition that the communicative situations are repetitive, both in terms of register of the utterances and of content of the communications, could be fairly reasonable.

In order to verify this hypothesis, a topic modeling on the English side of the English-Italian Europarl has been performed. As the choice of the topics is arbitrary, different customisations have been tried, finally deciding for a number of 20 topics with hyperparameter optimization. Results are shown in Table 1.

It is clearly possible to associate some of these clusters of words to specific topics. For example Topic 1 is related to energy production (and related environmental issues); 3 is about warfare (with a stress to the Iraq war); 4 immigration; 5 medical treatment of addictions; 6 food production; 7 Euro-economics; 8 global market; 9 Middle East; 12 transports; 13 finance; 17 human rights; 18 Eastern Europe; 19 primary sector; 20 legal aspects. The remaining groups of words appear to be quite similar, since they do not suggest very specific terminology and contain almost completely terms around the European parliament activities, widely used throughout the debates in the European parliament: *president*, *report*, *parliament*, *council* etc.

Taking a look at the distribution of the single document across the topics, it appears that very few documents show a remarkable unbalancedness towards the first topic they belong to, meaning that the affiliation of a single document to specific topics is weak. Probably this is due to the fact that sessions of the European parliament can deal not only with a single topic at time, and it is not the case where a corpus document always corresponds to a single argument. It can be concluded that, although this experiment has led to a better understand of the composition of Europarl (or at least of the portion of English Europarl aligned with its Italian counterpart), it has demonstrated that this corpus is too homogeneous to be used for the extraction of subsets for SMT.

2.2 Document similarity experiment

Taking into account the conclusion achieved in the previous section, an attempt of selecting a subset of training data for a specific translation task from Europarl has been made anyway. The aim is to pick up a selected portion of Europarl documents that are recognized to be the most similar to those to be translated, and use them as training data for SMT. Since the topic modeling has proved not to be very helpful for this purpose, the cosine similarity measure has been chosen as alternative to compute the distance between a test document to be machine-translated and all the documents in Europarl. The test document has been randomly selected from the Internet, it is a journal article about a controversy in the Catholic church, dated 10 September 2009; it was written in Italian and provided with English human translation (which will be used later as benchmark for MT evaluation); its size is around 1350 words per language on 47 sentence pairs.

In order to select and use the most similar data in Europarl to this document the following procedure has been followed:

Table 1: Topics in Europarl.

Topic	Keywords
1	energy climate european change emissions eu environmental gas nuclear industry policy renewable research europe countries environment states sources development
2	european europe president union mr people parliament citizens treaty political today presidency time eu constitution member states future world
3	international people peace situation war aid united union military european resolution president security government support humanitarian mr country iraq
4	european states member rights report data protection eu legal justice terrorism immigration law asylum citizens countries people union crime
5	health research people programme european diseases tobacco drugs states disease member human public europe information patients care framework treatment
6	directive food products health environment safety proposal animal protection animals amendments consumer legislation waste water environmental consumers commission substances
7	economic euro financial european growth monetary bank stability crisis economy policy central states market currency pact countries markets investment
8	countries trade development world eu european agreement union developing economic international africa wto china aid cooperation negotiations agreements global
9	israel education european palestinian cultural people culture programme sport israeli young peace languages middle support palestinians europe media training
10	council european union policy countries presidency mr president rights parliament political states office common agreement security enlargement summit process
11	commission mr president member european important time make commissioner made work states debate point parliament question council fact clear
12	transport safety european road air proposal tourism traffic rail report maritime directive europe sector mr sea environmental states passengers
13	budget commission parliament financial european funds committee programme policy year eur budgetary money report support council fund aid court
14	parliament mr vote report president committee amendment group european members procedure amendments house minutes rules mrs resolution voting rapporteur
15	report social women european policy eu states member employment people development work support voted europe economic writing union regions
16	report european union mr parliament policy council countries committee social economic europe community people president agreement treaty question employment
17	rights human people country president democracy political situation resolution government freedom china death democratic european eu mr respect elections
18	eu european union turkey russia countries accession country ukraine russian negotiations turkish political relations enlargement romania report cooperation region
19 8	fisheries fishing agricultural policy sector report production proposal farmers support agriculture rural market measures european aid regions commission reform
20	directive market services european proposal report member competition states parliament internal companies legal workers public rights legislation regulation protection

1. The cosine similarity has been computed between the test doc and each one of the 6,216 Europarl files (on the English side of the EN-IT subcorpus);

2. Files have been sorted according to the cosine similarity score and the first 500 have been selected;

3. The result of this selection (303,615 sentences pairs, around 7 million words per language) has been used to train a SMT system in Moses [8];

4. The obtained parameter file has been used to translate the test document (in the English>Italian direction);

5. The previous steps 3 and 4 have been repeated substituting the training data selected with cosine similarity with other 500 documents randomply extracted from Europarl, in order to obtain a term of comparison and validate the quality of the resulted translations;

6. MT evaluation of the quality of these translations has been carried out by using BLEU [12];

7. The whole above described process has been repeated in the opposite language direction (Italian>English).

The results of this analysis are presented in Table 2. They confirmed that the selection of a subset of documents that are more similar to the to-be-translated document is useful, giving - at least according to the automatic evaluation conducted by BLEU - a better translation than using random documents even in a corpus like Europarl that is circumscribed to few communicative situations. However, there are some considerations that need to be done about these results: although they are positive they are not extremely exciting, especially considering that the random training set was much smaller than the one selected with cosine similarity (117,973 sentences pairs and about 2 million words per language). This is most probably due to the fact that cosine similarity calculates the similarity between the two documents as a single whole feature, without distinguishing between the several aspects that make a texts similar to another one, like terminology, sentence structure, grammar, length etc. So there is the possibility that training data appropriate and useful in terms of genre or domain but with size is not good to test texts are discarded from this kind of selection.

To sum up, some strategies to better understand the composition of one of the most used parallel resources freely available to the MT community has been applied, including the possibility to select the most suitable data for a specific SMT task among all the documents contained in it. The conclusion is that, even if these strategies can be helpful for their usability on Europarl, they are quite limited for the intrinsic features of this corpus - that provides a huge quantity of data but without a very big assortment of textual varieties. Previous research about domain adaptation applied to SMT [9] has shown how some help could come from integrating a benchmark resource like Europarl with other data, obtained by retrieving parallel corpora from the web. Europarl itself could be considered as a webcorpus, since their authors have built it after having downloaded and aligned

Table 2: BLEU score for the cosine similarity experiment

Direction	Training set	BLEU score
IT>EN	500 most similar	27.5
IT>EN	500 random	26.1
EN>IT	500 most similar	26.5
EN>IT	500 random	23.3

the proceedings from the website of the European parliament[1], despite the fact that these texts were not originally created with the specific purpose of being published on webpages. In any case, in the next sections the way to explore this possibility is described, practically considering which instruments can be used to retrieve parallel documents from the web and understand their nature.

3. PARALLEL CORPORA FROM THE WEB

Crawling the web in order to find textual data is a common practice in corpus linguistics, and even if coming with a surplus of difficulties comparing to monolingual corpus creation (finding parallel pages or websites or pairing a webpage with their translated counterpart in another language are non-trivial tasks) also the creation of parallel corpora from the web was explored especially in the last decade. However, the literature shows that these technical difficulties led the researchers in the sector to focus more on the mechanisms needed to create parallel webcorpora rather than on an in-depth analysis of the results from the point of view of the kind of texts that is possible to find on the Internet. In the next few subsections an outline of the background studies on this topic will be given, followed by the way in which some of these strategies have been applied in the present research in order to understand the composition of a particular region of the parallel web.

3.1 Related works and known issues

The forerunner of using the web as a source for collecting parallel corpora is Philip Resnik and his sistem STRAND, described since the late nineties in a series of papers. In the latest, *The web as parallel corpus* [13], the core STRAND system is explained as well with several improvements comparing to previous versions. The main idea behind this approach is to find web pages that exhibit a parallel structure at the level of url and/or page composition, and that could be mutual translations; in the practice this is done relying on the performances of AltaVista advanced search engine options that permit to find pages containing hypertexts links to different language versions of the same document (parents) or pages that contains a link to a translation of their content in another language (siblings). The so-retrieved pages are then subject to a candidate pairs detection task that can be carried out with several strategies, like automatic language identification, url matching, document lengths and in the last version also a content-based similarity measure, to detect pairs of pages that do not present similarity just at the level of structure.

As shown with STRAND, the operation of grabbing parallel texts from the web is usually divided into two steps: loca-

[1] `http://www.europarl.europa.eu`

tion of websites that may have translated texts, and extraction and alignment of bitexts from these candidates. Other systems, developed independently from STRAND but employing similar approaches, have seen the birth in the same period, and some of them have explicated some useful (although not universally true) presuppositions that can be made when starting a collection of parallel texts from the web, like assuming that parallel texts usually are present in the same site [4] and that national top level domains are expected to have sites in the language of respective countries [10].

Several approaches in literature rely on the functionalities given by commercial search engines, like Altavista for STRAND or the application programming interfaces (APIs) to commercial search engines Google [11] or Yahoo! [1]. It is worth to point out -for reasons explained in the next sections - that the reliability on these systems, provided by famous search engines, comes with some disadvantages: there are intrinsic limitations with regards to the number of results per query and the unknown criteria about how documents are selected, their availability is for undefined periods of times after which the service may be no longer supported or available to users[2]. However, even if coming with these disadvantages the reliability on search engine functionalities is a possibility that gives undeniable advantages, considering the fact that it gives easy access to previously unknown parallel texts as the next section will explain in detail.

3.2 Corpus construction

The strategy here used to mine the web looking for previously unknown parallel pages is very similar to the original one described by Resnik: to make use of a search engine relying on its specific research functionalities. Resnik used Altavista, but the employment of these strategies has the drawback of relying on possibly discontinued services. In fact at the moment it is not possible to exactly replicate their specific algorithm because this search engine is no more available with the same options[3] of the time that the article was written.

However, it has been possible to re-implement a similar procedure using the search engine query algorithm that is part of the BootCaT toolkit, as just said currently relying on Bing. As seeds tuples the ones produced to build the large web corpus of English ukWaC [6] have been used, adding to each of these 1000 lines the use of two advanced operators: `site:` and `inanchor:`. The first is used to look for sites that fall under the intended national top level domain (in this case `.it`) and the second to find pages containing a specified term in the anchor text, in this case common English-related features of URLs (`en`, `eng`, `english`).

In practice, the search has been for pages in Italian websites that most likely contain English versions of their content.

[2]BootCaT [3] used Google APIs in its original version, but then moved to Yahoo! after Google started giving strong limitations to its service and now BootCaT relies on Bing APIs after Yahoo! discontinued the use of their SOAP APIs (see `http://bootcat.sslmit.unibo.it/wiki/doku.php?id=release_notes:frontend:0.60`).

[3]The shut down of Altavista by its owner Yahoo! began in May 2011.

Table 3: First 10 lines of the seeds list.

```
inanchor:en site:it grey gently
inanchor:en site:it drawing totally
inanchor:en site:it path eating
inanchor:en site:it watching explanation
inanchor:en site:it dealt lack
inanchor:en site:it radical organised
inanchor:en site:it relationships studied
inanchor:en site:it gets accused
inanchor:en site:it conservative hoping
inanchor:en site:it realise increasing
```

Queries are issued asking for 50 results per query (this is the maximum available from Bing APIs).

Table 4: Results for the first 3 queries on an English-Italian pair.

```
CURRENT_QUERY inanchor:en site:it grey gently
http://www.domusweb.it/en/architecture/teshima-art-museum-/
http://gilda.it/gandalf/italiano/giochi_di_ruolo/girsa_rolemaster/moduli/
amroth/amroth.htm
http://www.beppegrillo.it/en/politics/
CURRENT_QUERY inanchor:en site:it drawing totally
http://en.metals.it/productive-cycle-punching-c-104_133.html
http://flashandpartners.it/en/
http://www.dieproofs.it/english/prove_artista_eng.html
http://digilander.libero.it/cuoccimix/ENGLISH-automotorusse4(lada).htm
http://www.digicult.it/digimag/article.asp?id=1141
http://www.domusweb.it/en/architecture/post-carbon-loft-
http://architettura.it/artland/20020515/index_en.htm
http://en.museiincomuneroma.it/mostre_ed_eventi/mostre
http://www.beppegrillo.it/en/2010/06/
http://www.beppegrillo.it/en/information/
http://www.asianews.it/index.php?l=en&art=22711&size=A
http://www.domusweb.it/en/products/?idtema=5515?idtema=5530&inizio=25&da=1
http://www.domusweb.it/en/products/?idtema=5515?idtema=5562&inizio=13&da=1
http://www.disabilitaincifre.it/allegati/RECOMMENDATION_R(92)6.htm
http://www.beppegrillo.it/en/2009/08/
http://www.domusweb.it/en/products/?idtema=5515?idtema=5322&inizio=49&da=1
http://archivio.lanottebianca.it/nb2006/en_programma.html
http://www.pierpaoloricci.it/download/downloadsoftware_eng.htm
http://www.beppegrillo.it/en/politics/
http://archivio.lanottebianca.it/nb2005/en_programma.html
http://www.beppegrillo.it/eng/politics/
CURRENT_QUERY inanchor:en site:it path eating
http://www.guidatoscana.it/en/massa-carrara/visitare-massa.asp
http://www.visittrentino.it/en/localita/lavarone
http://www.ananda.it/en/courses/courses-meditation-and-self-realization
http://www.holly-wood.it/mlcad/install-en.html
http://www.visittrentino.it/en/vacanze_a_tema/neve/ski_area/dett/
ski-area-pampeago-predazzo-obereggen?areaId=A10
http://www.italia.it/en/discover-italy/emilia-romagna/ferrara.html
http://www.italia.it/en/discover-italy/tuscany/florence.html
http://www.beppegrillo.it/eng/ecology/
http://archivio.lanottebianca.it/nb2005/en_programma.html
http://www.beppegrillo.it/en/2009/08/
http://www.beppegrillo.it/eng/politics/
```

At the end three lists of urls for each language, sorted and unified in order to have one single list per language without repetitions have been produced. In order to extract those pages that actually are English translations of other Italian webpages on the same website, the final url list has been semi-automatically processed.

At this point such pages have been downloaded, and analysed by using Mallet. Results of this topic modeling are shown in table 5.

3.3 Analysis of results

One thing that need to be clarified now is that this is not a definitive description of the overall composition of the English-Italian portion of the Internet: the following analysis shows what is possible to intuitively retrieve using the search engine method and have an approximate idea of the kind of the indexed websites that represent a bilingual web space. However, since the accessibility to single site depends on their presence and positioning on the search engines, the

Table 5: Topics in the EN-IT corpus

Topic	Keywords
1	music art design work fashion world italian time film years style architecture de project works york great life la
2	di la engine il px car de brutale che pic version del mv agusta pics hp della news con
3	january inter milan time news photos derby don team ac great good people day back comments ll search italia
4	china peace arab christians vatican chinese world muslims christian government years year country case rsquo bishops samir mgr people
5	hotel di area sea wine day city km offers visit rooms room florence centre located italy free beautiful town
6	century di museum city church ancient san town built art roman del area rome st building palazzo history archaeological
7	italian international italy information university research di european law services public development students data company people financial system management
8	water high production system products quality made product time light energy range type materials oil process air design control
9	data time set file software information user version click web function download find system show distribution site number image
10	church god pope life world faith vatican people time Catholic holy benedict xvi christ great cardinal council human ii

present analysis shows what is actually possible to retrieve with this method; other existing strategies to find previously unknown multilingual websites [2] rely on the crawling of large web directories like the Open Directory Project[4], and again the possibility of benefiting from parallel sites is circumscribed only to those sites that are listed (in this case by human users to specific directories).

Considering the generated topic: most of the documents shows a strong belonging to the first topic they have been associated with, so we can graphically generate a plausible representation of the distribution of each topic in the whole context of the corpus.

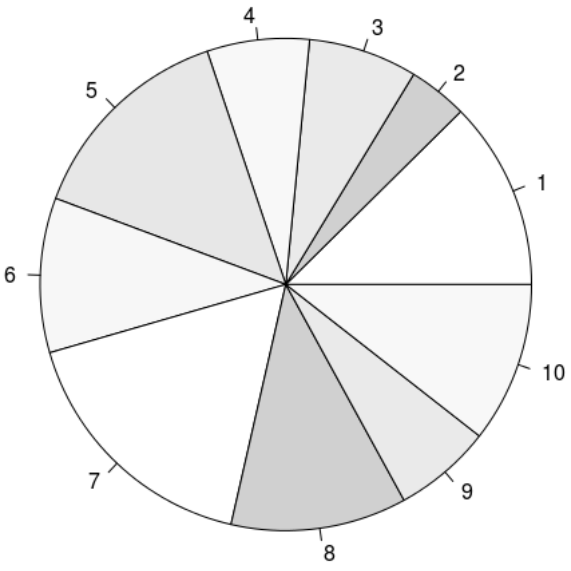

Figure 1: Distribution of the topics

[4] `http://www.dmoz.org/`

A sample of 10 randomly-selected documents for each of these topic have been selected and read by the author in order to understand the overall nature of the single portions of the corpus associated to each group of keywords.

1. The first topic suggests pages containing text related to fashion, art exibitions, recent trends in architecture, cinema, music etc. In fact it appears to be composed by articles coming from (but not exclusively) the online version a famous fashion and lifestyle magazine.
2. The second topic appears to be instead more problematic to be understood, since it contains several stopwords in Italian, such as article and prepositions; it is worth to say that English stopwords have been stripped out when preparing the corpus for topic modeling, since this is actually an English corpus. The generation of this topic can suggest the deletion of the original language counterpart's stopwords as well, but this can be risky since it would modify several compound proper nouns contained in the corpus as it has been revealed in the case. This portion of the corpus collects pages mainly from two sites: a fan site about vintage cars produced in the Soviet union, and a motorcycle manufacturer site; the unknown words *agusta* e *brutale* have reference to proper nouns of brands.
3. The third topic contains words related to football, in fact great part of this portion is made by news about football matches. However these keywords could be misleading since this part of the corpus also contains news about only other sports and other kinds of entertainment, such as tv programs and videogames. Words like *comments*, *back*, *search* are due to the presence of recurring boilerplate in the football news, that - again - almost completely have reference to a single website.
4. Also the documents collected around the fourth topic have reference to a single website, in this case a news website promoted by a Catholic missionary group (that generated a list of religious-related terms).

5. The fifth topic instead collects contributions from many single sites. Clearly suggested by the keywords, the manual analysis confirmed the presence of English versions of webistes of tourist accommodations such as hotels, residences, bed & breakfast etc.

6. The sixth topic is related to texts about cultural tourism and history of art. The samples have shown that websites of artistic attractions and personal pages with more pedagogical purposes belong to this category.

7. The seventh topic is less homogeneous that the previous ones. Contributes to this large part of the corpus come from presentation of university programmes and related bureaucratic guidelines, but also legal disclaimers from private companies and translations of legal statuses.

8. This is another topic that collects contributions from many sources but present a quite specific genre of texts: companies presenting their products. Words like *high*, *quality*, *production*, *range* suggests the intention of promoting themselves, and this was confirmed by the sample data. The little portion manually analysed revealed the presence of several kind of companies, from the primary sector to services; but the biggest part appeared to be related to mechanical and electrical product engineering.

9. The ninth topic appears to be about computing. Around this topic are collected webpages of various origins, from academic workshops to web tutorial to description of programs, plugins etc.

10. As suggested by the keywords, the tenth topic collects articles around it articles about the Catholic church, in particular news and debates. Most of the documents have reference to two online magazines, but in this portion of the corpus there are also pages coming from other sources and not necessarily related to the Catholic church hot topics, since they discuss about philosophy, sacred music or other religions.[5]

This analysis has provided an overview of the downloaded documents, and the possible composition of the so generated corpus. The topic modeling technique revealed the major presence of documents taken from recurring websites, in a way that several topics corresponds to particular webistes and their content. Even if this circumscribes the possibilities of having variety in the corpus, it can give advantages for the problem of retrieving document pairs, since it allows to develop a specific strategy for these large websites basing on how they organize their content. On the other hand, there are groups of different websites but having similar purposes, like showcasing products and services or displaying guidelines and rules. In this case there are recurring language structures spread across pages coming from different sources, that could be useful e.g. for generalizations about grammar or terminology.

[5]The test article described in section 2.2 was taken from this collection.

4. FURTHER WORKS AND CONCLUSIONS

Since the purpose of this paper is only to describe a strategy to explore and understand the composition of a bilingual web space, only an aspect of the building of a parallel corpus from the web (location of webpages that may contain parallel texts) has been taken into account, i.e. the ability to find pages that contain translated text[6]. This is just the first step, as it needs to be followed by further fundamental operations, above all the location of the counterparts in other(s) language(s) of the previously obtained webpages and the sentence alignment of their content.

This study is part of a major PhD project aimed at exploring how much the successfulness of SMT technology depends on the exploitation parallel corpora basing on a good match between the text typology (genre and domain) of training data and to-be-translated texts. This means that the preliminary study here presented is going to be continued with the creation of some parallel corpora from the web following the strategy described in section 3.2, and the mentioned further steps. Previous literature has shown that the generation of candidate pairs can be performed via a series of heuristics such as url substitution rules, analysis of document lengths and structural filtering. Since there is not an established state-of-the-art system to perform this job and very few tools able to do that are freely available[7] - some of them are currently experimented, and possibly integrated in a single framework that would be able to cover the whole chain, from the search for parallel pages to the extraction of parallel pairs and their alignment. Another thing that has not been mentioned in this paper but that can be very useful to expand the size of a parallel corpus is that the experiment here described deals with single pages rather than whole websites: even if the search engine selected only a webpage from a site, there is the real possibility to find more parallel texts, going up to the main website and exploring the tree of its contents, collecting all the others parallel pages contained in it. To conclude, this was an example focused on a particular language pair (Italian-English), but it would be interesting to explore different language pairs in order to understand what different kind of parallel data are on the web depending on particular language pairs.

To sum up, in this paper some studies about the importance of the analysis of parallel corpora have been conducted, starting from the considerations that not only it is advisable to have knowledge of the data used (in particular, in the second part of the paper, the exploration of the composition of a parallel corpus retrieved from the web has been attempted), but also that this can be helpful in order to select the most suitable textual data for our specific purposes. The strategies here proposed can have wide application for everybody who wants to better benefit from existing or to-be-constructed parallel resources, from the creation of translation memories for computer-assisted translation to the creation of larger parallel corpora. The next step will be that of exploring this last possibility, since there are not standards about the dimensions of parallel corpora from the web and their size can remarkably change among different language

[6]Scripts and commands employed in the experiments here described can be found at `http://smlc09.leeds.ac.uk/marco/tools.html`.

[7]A rare example is Bitextor [5].

pairs.

5. ACKNOWLEDGMENTS

The research described in this paper is supported by FP7 ICT grant
ACCURAT Grant agreement no.: 248347

6. REFERENCES

[1] J. J. a. Almeida and A. Simões. Automatic Parallel Corpora and Bilingual Terminology extraction from Parallel WebSites. 2010.

[2] L. Barbosa, S. Bangalore, and V. K. S. Rangarajan. *Crawling Back and Forth: Using Back and Out Links to Locate Bilingual Sites*. 2011.

[3] M. Baroni and S. Bernardini. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the LREC 2004 conference*, volume 4, pages 1313–1316. ELRA, 2004.

[4] J. Chen and J.-Y. Nie. *Parallel Web text mining for cross-language information retrieval*, pages 62–77. Paris, 2000.

[5] M. Esplà-Gomis and M. L. Forcada. Combining content-based and URL-based heuristics to harvest aligned bitexts from multilingual sites with Bitextor. In *Fourth Machine Translation Marathon Open Source Tools for Machine Translation*, 2010.

[6] A. Ferraresi. Building a very large corpus of English obtained by Web crawling: ukWaC, 2007.

[7] P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, 2005. AAMT.

[8] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[9] P. Koehn and J. Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[10] X. Ma and M. Liberman. *Bits: A method for bilingual text search over the Web*. 1999.

[11] M. Mohler and R. Mihalcea. Babylon Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages. In *LREC'08*, 2008.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[13] P. Resnik and N. A. Smith. The Web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, Sept. 2003.

[14] M. Steyvers and T. Griffiths. Probabilistic Topic Models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum, 2006.