# Parallel Global Voices: a Collection of Multilingual Corpora with Citizen Media Stories

**Prokopis Prokopidis, Vassilis Papavassiliou and Stelios Piperidis**
Institute for Language and Speech Processing
Athena Research and Innovation Center, Athens, Greece
{prokopis, vpapa, spip}@ilsp.gr

**Abstract**

We present a new collection of multilingual corpora automatically created from the content available in the Global Voices websites, where volunteers have been posting and translating citizen media stories since 2004. We describe how we crawled and processed this content to generate parallel resources comprising 302.6K document pairs and 8.36M segment alignments in 756 language pairs. For some language pairs, the segment alignments in this resource are the first open examples of their kind. In an initial use of this resource, we discuss how a set of document pair detection algorithms performs on the Greek-English corpus.

**Keywords:** parallel corpus, independent news stories, document pair detection

## 1. Introduction

This work describes Parallel Global Voices (PGV), a collection of parallel corpora created from the Global Voices[1] group of websites. According to the Global Voices (GV) project description, GV "is a border-less, largely volunteer community of more than 1400 writers, analysts, online media experts and translators". The GV community has been reporting since 2004 on trending issues and stories published on social media and independent blogs in 167 countries.

In this paper we provide an overview of how we used this content to automatically generate an open set of parallel corpora that exhibit some interesting features as far as topics and language pairs are concerned. The rest of the paper is organized as follows. After discussing related work in Section 2, we report in Section 3 on how the Global Voices content was crawled, processed and aligned at document and sentence level. Section 4 provides details on the size and characteristics of monolingual and parallel sub-corpora in PGV and on format and availability. As an example of the potential use of the parallel corpora presented here, we use them to examine methods for the detection of parallel web pages and discuss results in Section 5.

## 2. Related Work

Parallel corpora like PGV are important for a number of applications (Tiedemann, 2011) including SMT, induction of bilingual lexica and contrastive studies of language use. Tlaxcala (Toral, 2014) was the first publicly available collection of parallel and monolingual corpora acquired from independent news sources. The largest parallel corpus in this 15-language resource is English-Spanish with 66.8K sentence pairs. Rettinger et al. (2014) compiled a parallel corpus of 300 English/Spanish/German GV articles, which they hand-annotated with semantic groundings of named entities and concepts to cross-lingual linked data extracted from Wikipedia. Chahuneau et al. (2013) used an English-Swahili parallel corpus obtained by crawling GV and reported significant improvements in translation quality when translating to Swahili. Finally, a GV Malagasy-English parallel corpus, collected and aligned at sentence level by V. Chahuneau, is available from `http://www.ark.cs.cmu.edu/global-voices/`. This work is to the best of our knowledge the first that extracts parallel and monolingual resources for all languages and language pairs in the GV websites. The datasets comprising this resource are smaller than the ones obtained from mining large scale crawls (e.g. Smith et al. (2013)). In contrast, this work is a focused effort to extract highly parallel documents by exploiting the well-defined structure of a multilingual site.

In another research line, Harlow and Johnson (2011) discuss how the 2011 Egyptian protests were depicted in, among other sources, 66 stories from a major US newspaper and 49 documents on the English GV site. The datasets presented in this paper could make similar comparisons easier, even from a multilingual perspective.

## 3. Acquisition and Processing

The content of the GV websites was crawled in July-August 2015 and in January 2016 by the authors. The crawl resulted in 174.63K documents in 41 languages (with traditional and simplified Chinese counting as two different languages). After crawling, we exported each document's content to XCES-compatible XML files. At this stage, we took advantage of the fairly homogeneous HTML structure of the crawled web documents to identify the actual content, remove boilerplate, and export text segmented into paragraphs. As a result, we obtained a total of 2.50M paragraphs, of which approximately 293K were annotated in the exported files as in a language other than the main language of the document. This annotation was predominately based on the largely consistent markup of related text chunks in the original HTML documents. For the rest of the content, we relied on the results of a language identifier[2] to similarly annotate certain types of obvious errors in the authors' markup like, for example, a paragraph with Chinese text in an English document.

[1] `https://globalvoices.org/`

[2] `https://github.com/shuyo/language-detection`

During this stage, we also extracted and stored in the XML files metadata information regarding publication date, authors and translators. The year with the largest number of published documents (according to the number of files crawled) was 2011 with 25.8K posts for all languages.
The metadata set in the exported files also included information on the language-dependent topic and region key terms with which authors and translators tagged their documents (Fig. 1). We observed that each document may be tagged with more than one key term. The temporal evolution of certain key terms' frequency (cf. Fig. 2) illustrates the interest of content creators in posting content about breaking events. As in Smith et al. (2013), we applied Latent Dirichlet Allocation (Blei et al., 2003) to explore the topics of the crawled dataset. In Table 1 we select 10 of the 20 topics generated from the 61.5K documents of the English corpus using the Mallet toolkit (McCallum, 2002). The top representative tokens for the selected topics reflect the interest of the content creators in, among other things, politics and elections (1), civil, sexual and socio-economic rights (2), disasters and the environment (3), demonstrations and police reaction (4), labour (5), specific geographic regions (6-8), organization of the GV network (9) and culture and online media (10). In a similar experiment with a lemmatized and unaccented version of the much smaller Greek dataset (3.6K documents), we observed interest for similar topics: topics 1-8 seem quite similar to their English counterparts with the same id.

```
<keyTerm>Freedom of Speech</keyTerm>
<keyTerm>Human Rights</keyTerm>
<keyTerm>Refugees</keyTerm>
<keyTerm>Technology</keyTerm>
<keyTerm>War & Conflict</keyTerm>
<keyTerm type="region">Germany</keyTerm>
<keyTerm type="region">Hungary</keyTerm>
<keyTerm type="region">Syria</keyTerm>
```

Figure 1: Key terms exported from a September 2015 English document on the refugee crisis and the reactions it generates among social media users across Europe

Most importantly for generating the parallel resources, at the exporting stage, we parsed the links of each document and stored information about its translation counterparts, thus creating a set of document pairs. We then used the language dependent sentence splitters included in the Morphadorner NLP suite (Burns, 2013) to split paragraphs in the XML files into sentences. Segment alignments (with alignments of up to 1:2 and 2:1 sentences) were then extracted from each document pair with the Maligna sentence aligner (Jassem and Lipski, 2008) using default values and without any adaptation of the aligner to the language pair under examination. Paragraphs that were annotated as being in a language other from the main document language were excluded from the alignment process.

## 4. Resource Description

Most of the crawled documents (151K, 86.51%) are involved in at least one document pair. English is the only language with a relatively large (21.8K, 35.39%) percentage of documents for which a translation in another language does not exist. Table 2 presents the set of languages involved, the 3-letter language codes used and basic size information for those documents participating in translation pairs, and the number of their paragraphs. As Fig. 3 shows, English is the only language with a larger number of source documents (90.65% of total documents). For Spanish, the second largest language in the resource, only 7.37% of the texts are source documents. We detect whether a document is a translation based on whether it includes metadata for a translator. We did not find a way to reliably identify the source language from which a specific translation was created. However, we observed that each source document is translated (potentially via a pivot document) into 2.77 languages on average, thus generating combinations of sentence alignments as in the examples of Table 3.
Overall, the parallel resource comprises 302,617 document pairs and 8,356,943 segment alignments for 756 language pairs, with 27.62 segment alignments per document pair on average. The information on segment alignments is calculated after filtering out circa 427.4K 0:1 and 1:0 cases. The distribution of the number of segment alignments on all document pairs (Fig. 4) reflects the fact that a substantial part of original and translated content contains short descriptions ("quick reads" in the GV Posting Guide[3] terminology) of online content external to GV. The main part of the rest of the content includes longer stories that follow a distribution which, as far as size in segment alignments is considered, seems to be in agreement with the GV guideline for longer articles of 500-1000 words length. Language pairs involving combinations of each of eng, fra, ita, mlg,

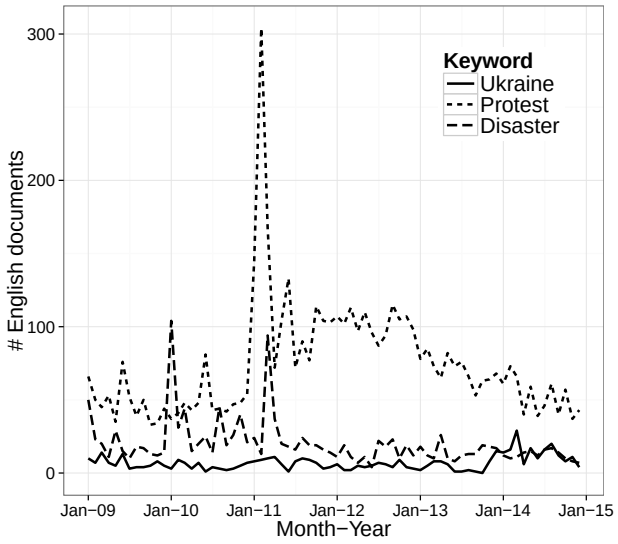


Figure 2: Three key terms used for tagging English documents. Peaks correspond to the 2014 Ukrainian crisis (Ukraine), the 2011 beginning of the Arab Spring (Protest), the 2010 Haiti earthquake and the 2011 Japan earthquake & tsunami (Disaster)

[3] https://community.globalvoices.org/guide/technical-guides/gv-posting-guide/

| id | Top representative tokens |
|---|---|
| 1 | election political party president government vote country minister people candidate opposition state leader power |
| 2 | woman child student school education university rights girl health man law women gay family sexual |
| 3 | city people water area photo report flood earthquake build road village local disaster day project |
| 4 | police protest rights government people arrest human report activist group video case force court march |
| 5 | government country company money year work price worker economic million pay people market make increase |
| 6 | egypt egyptian saudi bahrain blogger arab post tweet morocco write tunisia twitter libya arabia revolution |
| 7 | china chinese hong kong people taiwan myanmar beijing government netizen weibo media news official mainland |
| 8 | israel gaza israeli war syria lebanon iraq syrian refugee palestinian lebanese attack kill bomb palestine |
| 9 | blog project global language video voices work community world media people film event online social |
| 10 | media internet information online user blog website social facebook twitter news blogger report site government |
| 1 | πολιτικος κομμα προεδρος εκλογη χωρα δημοκρατια νεος κυβερνηση εθνικος λαος πολιτης εξουσια πρωθυπουργος |
| 2 | γυναικα κοριτσι αντρας σεξουαλικος βια βιασμος γαμος δικαιωμα θυμα ομοφυλοφιλος ινδια φυλο κοινωνια |
| 3 | περιοχη πολη καταστροφη νερο κατοικος φωτογραφια ενεργεια σεισμος ιαπωνια χωριο πυρηνικος ποταμος |
| 4 | διαδηλωση διαδηλωτης βιντεο διαμαρτυρια αστυνομια φωτογραφια αστυνομικος πορεια πλατεια πολη κινημα |
| 5 | δικαιωμα νομος κυβερνηση ελευθερια ανθρωπινος χωρα ιστοσελιδα αρχη πολιτης διαδικτυο υπηρεσια ασφαλεια |
| 6 | αιγυπτος υεμενη επαναστση σαουδικος αραβια τυνησια αιγυπτιος twitter γραφω λιβυη αραβικος αιγυπτιακος |
| 7 | κινα κινεζικος κονγκ χονγκ κινεζος κορεα weibo βορειος φαγητο ιαπωνια πεκινο china τοπικος πιατο χρηστης |
| 8 | ελλαδα γαζα προσφυγας ισραηλ ελληνικος παλαιστινιος ισραηλινος μεταναστης ευρωπαικος ελληνας παλαιστινη |
| 9 | συρια συριος λιβανος συριακος syria ασαντ καθεστως δαμασκος μιανμαρ ταιλανδη λανκα σρι επαναστση |
| 10 | φωτογραφια πολη ανθρωπος τεχνη σελιδα καλλιτεχνης εικονα αδεια χωρα φωτογραφος βιντεο δρομος παραδοσιακος |

Table 1: Ten English and ten Greek topics and their representative tokens

| Code | Language | Documents | Paragraphs | Code | Language | Documents | Paragraphs |
|---|---|---|---|---|---|---|---|
| amh | Amharic | 41 | 917 | khm | Khmer | 32 | 575 |
| ara | Arabic | 3560 | 46570 | kor | Korean | 350 | 7583 |
| aym | Aymara | 679 | 10368 | mkd | Macedonian | 2249 | 37053 |
| ben | Bangla | 7348 | 113333 | mlg | Malagasy | 9200 | 181388 |
| bul | Bulgarian | 288 | 4417 | mya | Burmese | 106 | 1519 |
| cat | Catalan | 727 | 16648 | nld | Dutch | 1273 | 24424 |
| ces | Czech | 464 | 9168 | ori | Odia | 20 | 264 |
| dan | Danish | 316 | 8327 | pol | Polish | 1606 | 35703 |
| deu | German | 2608 | 47280 | por | Portuguese | 5128 | 80320 |
| ell | Greek | 3623 | 43270 | rum | Romanian | 69 | 1765 |
| eng | English | 39743 | 524103 | rus | Russian | 3530 | 67970 |
| epo | Esperanto | 144 | 1673 | spa | Spanish | 30425 | 451721 |
| fas | Farsi | 760 | 7273 | sqi | Albanian | 293 | 4208 |
| fil | Filipino | 254 | 3587 | srp | Serbian | 978 | 21428 |
| fra | French | 15422 | 243926 | swa | Swahili | 1405 | 20590 |
| heb | Hebrew | 22 | 469 | swe | Swedish | 339 | 8542 |
| hin | Hindi | 164 | 1158 | tur | Turkish | 153 | 3350 |
| hun | Hungarian | 515 | 9169 | urd | Urdu | 136 | 2604 |
| ind | Indonesian | 505 | 10084 | zhs | Chinese-simplified | 4996 | 88301 |
| ita | Italian | 4700 | 95133 | zht | Chinese-traditional | 5014 | 88861 |
| jpn | Japanese | 1886 | 39961 | - | Total | 151,071 | 2,365,003 |

Table 2: Number of documents (participating in translation pairs) and paragraphs for each language in the PGV

spa, zhs and zht with all other languages contribute 87.18% of the segment alignments. The largest language pair is eng-spa with 29.6K/724.8K document pairs/segment alignments, followed by eng-fra and fra-spa (Table 4).

An examination of the resource has led us to the conclusion that the overwhelming majority of the document pairs consists of parallel documents, i.e. that the percentage of links between comparable documents is negligible. Concerning another aspect of the resource related to alignment quality, we also believe that a large percentage of zero to one alignments in a document pair is a potential indication of exporting, sentence splitting and/or alignment problems, especially in pairs where one non-latin language (e.g. ben and zh[st]) is involved.

| | | |
|---|---|---|
| **eng**: His family were all still in Syria and he didn't want to leave them or his friends. | **ell**: Η οικογένειά του ήταν όλη ακόμα στη Συρία και δεν ήθελε να αφήσει αυτή ή τους φίλους του. | **ara**: ولايريد سوريا في تزال لا كلها عائلته كانت أصدقائه. يترك أو يتركهم أن |
| **eng**: In Bolivia, where unions are extensively formed by members of society, another group of workers have unionized: children. | **aym**: Bolivia markana, sindicatunakax markachirinakamp chikancht'atawa, yaqha qutux wawanakawa sintikalisasipxatayna. | **spa**: En Bolivia, donde los sindicatos están muy extendidos entre los miembros de la sociedad, otro grupo de trabajadores se ha sindicalizado: los niños [en]. |
| **eng**: Angola is a plurilingual country [en], with six African languages recognised as national languages as well as Portuguese as the official language. | **por**: Angola é um país plurilíngue, com seis línguas africanas reconhecidas como nacionais a par do português enquanto língua oficial. | **fra**: L'Angola est un pays multilingue [français], avec six langues africaines reconnues comme langues nationales ainsi que le Portugais qui est la langue officielle. |
| **eng**: But there is no solid evidence that such a planet exists. | **rus**: Но нет никаких убедительных доказательств, что такие планеты действительно существуют. | **zhs**: 但目前仍未有明确资料证明这样的星球存在。 |

Table 3: Examples of alignments in different languages and topics (with War & Conflict/Labor/Language/Science used as key terms in the English documents)

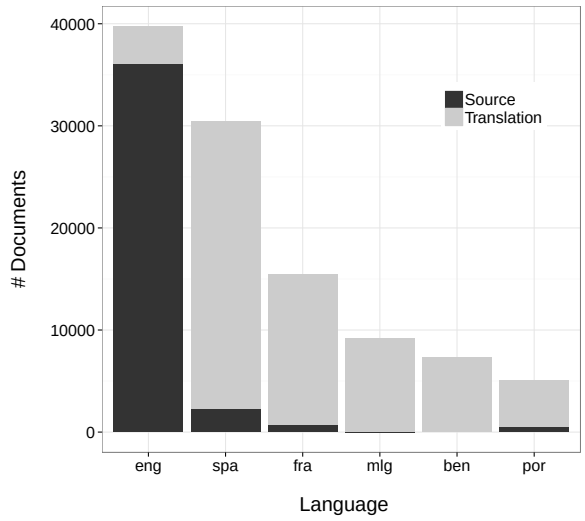


Figure 3: Source and translation documents in the 6 largest monolingual datasets

| Lang. Pair | Doc. Pairs | Seg. Alignments |
|---|---|---|
| eng-spa | 29,645 | 724,800 |
| eng-fra | 14,930 | 393,686 |
| fra-spa | 11,366 | 338,114 |
| eng-mlg | 8,893 | 262,177 |
| mlg-spa | 7,607 | 240,186 |
| ben-eng | 7,294 | 170,696 |
| fra-mlg | 4,805 | 154,117 |
| eng-ita | 4,609 | 152,887 |
| ben-spa | 5,524 | 144,091 |
| zhs-zht | 4957 | 134361 |

Table 4: The 10 language pairs with the highest number of segment alignments

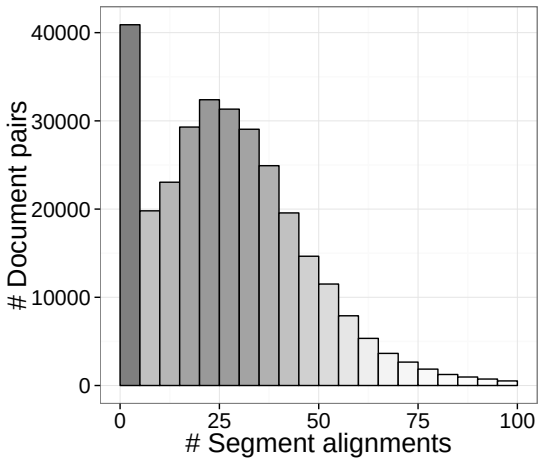


Figure 4: Distribution of the number of segment alignments on all document pairs

The fact that boilerplate detection and paragraph-level language identification are predominately based on the HTML structure could also affect resource quality. Taking into account that alignments suffering from noise related to these issues might be of limited or no use for downstream tasks (including training SMT systems), we counted, for the top language pairs, the number of (almost) identical alignments ($l1_1 \approx l1_2$ *and* $l2_1 \approx l2_2$) and of alignments in which the segment in l1 is identical to the segment in the l2 ($l1_1 \approx l2_1$). We observed that on average such alignments comprise less than 2.4% of these datasets. We did not include in these counts identical segments in l1 that have been aligned with different segments in l2, as in the examples of alternative translations in Fig. 5.

The original content from the Global Voices websites is available by the authors and publishers under a Creative Commons Attribution-Only license. The current version of the derivative resources described in this paper (i.e. the results of the automatic alignment at document and segment level) is distributed under the same Creative Commons li-

```
<tuv xml:lang="eng">
        <seg>What do you think?</seg>
 </tuv>
<tuv xml:lang="spa">
        <seg>¿Qué piensan ustedes?</seg>
        <seg>¿Qué creen ustedes?</seg>
        <seg>¿Qué piensas?</seg>
        <seg>¿Qué piensan?</seg>
        <seg>¿Qué opináis vosotros?</seg>
        <seg>¿Ustedes qué creen?</seg>
        <seg>¿Qué opinas?</seg>
        <seg>¿Cómo lo veis?</seg>
 </tuv>
```

Figure 5: Alternative Spanish translations for an English sentence

cense from `http://nlp.ilsp.gr/pgv/`. This web page includes links from where non-0:1|1:0 aligned segments for each language pair can be downloaded as one TMX file. Alternatively, segment-aligned versions of document pairs for each language pair can be downloaded and examined independently. Finally, compressed archives of the monolingual corpora in the XML format mentioned above are also available and, together with the list of document pairs, they can be used for further document and sentence alignment experiments.

## 5. Document Pair Detection Experiment

As mentioned in Section 3 the document pairs were detected by exploiting the web site graph and identifying pairs of web pages that are connected with specific links which denote that one web page is the translation of the other. Since this is not the case in many multilingual web sites, several methods have been proposed for extracting parallel content from multilingual websites. To this end, we carried out an experiment of examining methods that are integrated into ILSP-FC (Papavassiliou et al., 2013), an open source focused crawler for automatic acquisition of domain-specific monolingual and bilingual corpora from the web.

Our aim was to test methods that are language-independent and do not take advantage of specific properties of the Global Voices website. The current version of the Pair Detector module of the crawler does not use any language resources such as bilingual dictionaries or generated translations by MT engines (as in Barbosa et al. (2012)) to mine parallel webpages. In addition, during the experiment described here, we ommited a tool's method that exploits special patterns in URLs and links that point to candidate translations. Thus, we only used methods that are based on a) cooccurrences of images with the same filename in HTML source, b) edit distance of sequences of digits in the main content of webpages and c) structural similarity.

We evaluated these methods in the task of reconstructing the English-Greek parallel collection, that is of identifying the 3581 document pairs of this language pair. The recall and precision rates were 68.56% and 92.50% respectively. Even though the precision could be considered high enough for providing data to train an SMT system, the recall seems poor. Given that the GV site includes two types of documents (i.e. "quick reads" and longer articles), we examined how these metrics are affected by the length of document pairs (in terms of total words in the main content of both documents in a pair). Thus, we counted the number of real, detected, and correctly detected pairs (*GT*, *Detected*, and *Correct* columns in Table 5 respectively) with a length higher than the number in column *Pair Len*. In addition, we calculated recall, presicion and F-measure for each case.

As expected, the module performs better for long pairs since it is more likely that such web pages contain images and digits and that their text is split into more than one paragraphs. On the other side, it is hard to identify pairs of very short documents as shown in the last four rows of Table 5 for which the total length of a pair is less than 200 words. For instance, we present in Fig. 6 three documents, each consisting of two paragraphs: the module wrongly predicted a pair between documents *a* and *c* where `2012` appears the same number of times. However, it is worth mentioning that the main effect in the tool's performance concerns recall ($< 87\%$) while precision remains high.

## 6. Conclusions

We presented Parallel Global Voices, a new and open parallel resource comprising 302.6K document pairs and 8.36M segment alignments that were automatically generated for 756 language pairs, based on content provided by volunteers contributing to the Global Voices effort. Although most segment alignments concern only a few of these language pairs, we think that PGV presents several interesting features including the domains covered and the fact that, for certain pairs (e.g. ell-zh[st]), similar open resources are not available. For other pairs we improve the current situation with new resources. In future work, we aim to further augment the resource with newly published content and to exploit it in experiments involving comparison of sentence alignment algorithms, induction of bilingual lexica, SMT domain adaptation and cross-lingual annotation projection.

## Acknowledgements

This work was supported by the Abu-MaTran (FP7-People-IAPP, Grant 324414) project and the European Language Resource Coordination effort, CEF Programme. We would like to thank Sokratis Sofianopoulos for his help with sentence alignment.

## 7. Bibliographical References

Barbosa, L., Sridhar, V. K. R., Yarmohammadi, M., and Bangalore, S. (2012). Harvesting parallel text in multiple languages with limited supervision. In *COLING*, pages 201–214.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Burns, P. R. (2013). Morphadorner v2: A java library for the morphological adornment of english language texts.

Chahuneau, V., Schlinger, E., Smith, N. A., and Dyer, C. (2013). Translating into morphologically rich languages with synthetic phrases. In *Proc. of EMNLP*.

Harlow, S. and Johnson, T. (2011). Overthrowing the Protest Paradigm? How The New York Times, Global

| Pair Len. | GT | Detected | Correct | Recall | Precision | F |
|---|---|---|---|---|---|---|
| 1500 | 554 | 536 | 533 | 96,21% | 99,44% | 97,80% |
| 1000 | 1011 | 984 | 973 | 96,24% | 98,88% | 97,54% |
| 900 | 1110 | 1079 | 1067 | 96,13% | 98,89% | 97,49% |
| 800 | 1185 | 1150 | 1138 | 96,03% | 98,96% | 97,47% |
| 700 | 1273 | 1235 | 1222 | 95,99% | 98,95% | 97,45% |
| 600 | 1367 | 1329 | 1314 | 96,12% | 98,87% | 97,48% |
| 500 | 1442 | 1404 | 1384 | 95,98% | 98,58% | 97,26% |
| 400 | 1524 | 1480 | 1456 | 95,54% | 98,38% | 96,94% |
| 300 | 1644 | 1591 | 1556 | 94,65% | 97,80% | 96,20% |
| 200 | 1854 | 1771 | 1710 | 92,23% | 96,56% | 94,34% |
| 150 | 2198 | 2007 | 1910 | 86,90% | 95,17% | 90,84% |
| 100 | 2890 | 2399 | 2243 | 77,61% | 93,50% | 84,82% |
| 50 | 3514 | 2643 | 2445 | 69,58% | 92,51% | 79,42% |
| 0 | 3581 | 2654 | 2455 | 68,56% | 92,50% | 78,75% |

Table 5: Evaluation results on document pair detection

```
<body>
<p id="p1" type="title">Λάος:
    Συμμετοχή στους
    Ολυμπιακούς του 2012</p>
<p id="p2">Το Λάος έστειλε [
    en] τρεις αθλητές στους
    Ολυμπιακούς Αγώνες του
    Λονδίνου. Μια εργαζόμενη
    στο Παγκόσμιο Πρόγραμμα
    Διατροφής του ΟΗΕ στο
    Λάος επίσης εκπροσώπησε
    τη χώρα, όταν κλήθηκε να
    γίνει Λαμπαδηδρόμος
    νωρίτερα μες στο μήνα.
</p>
</body>
```

(a) gv-ell-20120731-12924.xml

```
<body>
<p id="p1" type="title">Laos:
     Participation in 2012
    Olympics</p>
<p id="p2">Laos sent three
    athletes to the 2012
    London Olympics. An
    employee of the United
    Nations World Food
    Programme in Laos also
    represented the country
    when she was invited to
    become an Olympic
    torchbearer early this
    month
</p>
</body>
```

(b) gv-eng-20120729-342659.xml

```
<body>
<p id="p1" type="title">Less
    Censorship in Thailand?</
    p>
<p id="p2">Jon Russell
    reviews the latest Google
     Transparency Report for
    the period of January to
    June 2012 and notes that
    there were fewer requests
     made by the Thailand
    government to censor
    websites that insult the
    monarchy.
</p>
</body>
```

(c) gv-eng-20121115-373025.xml

Figure 6: Three short documents (a and b are ell-eng translations) that can confuse document pair detection methods based on document structure and number co-occurence similarity only


Voices and Twitter Covered the Egyptian Revolution. *International Journal of Communication*, 5(0).

Jassem, K. and Lipski, J. (2008). A new tool for the bilingual text aligning at the sentence level. In *Proc. of Intelligent Information Systems Conference*, Zakopane, Poland.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. `http://www.cs.umass.edu/~mccallum/mallet`.

Papavassiliou, V., Prokopidis, P., and Thurmair, G. (2013). A modular open-source focused crawler for mining monolingual and bilingual corpora from the web. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 43–51, Sofia, Bulgaria.

Rettinger, A., Zhang, L., Berović, D., Merkler, D., Srebačić, M., and Tadić, M. (2014). RECSA: resource for evaluating cross-lingual semantic annotation. In *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. (2013). Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st ACL*, pages 1374–1383, Sofia, Bulgaria.

Tiedemann, J. (2011). *Bitext alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Toral, A. (2014). TLAXCALA: a multilingual corpus of independent news. In *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.