# TLAXCALA: a multilingual corpus of independent news

**Antonio Toral**
School of Computing
Dublin City University
Ireland
atoral@computing.dcu.ie

### Abstract

We acquire corpora from the domain of independent news from the Tlaxcala website. We build monolingual corpora for 15 languages and parallel corpora for all the combinations of those 15 languages. These corpora include languages for which only very limited such resources exist (e.g. Tamazight). We present the acquisition process in detail and we also present detailed statistics of the produced corpora, concerning mainly quantitative dimensions such as the size of the corpora per language (for the monolingual corpora) and per language pair (for the parallel corpora). To the best of our knowledge, these are the first publicly available parallel and monolingual corpora for the domain of independent news. We also create models for unsupervised sentence splitting for all the languages of the study.

**Keywords:** monolingual corpus, parallel corpus, under-resourced languages

## 1. Introduction

Parallel corpora constitute very useful language resources in multilingual applications of computational linguistics. For example, they are the backbone resources in the thriving field of statistical machine translation. However, the availability of these resources is limited and varies from language to language; while there are language pairs for which parallel corpora comprising millions of sentence pairs are available (e.g. English–Chinese and English–Arabic), for most language pairs the amount of parallel data available is rather limited, if any such resources exist at all.

The news domain can be considered a reasonably well represented one in terms of availability of corpora. However, to the best of our knowledge, all of them correspond to either corporate or government sources. Examples of sources of news corpora include Wall Street Journal[1] for English, Associated Press[2] for French, German and Portuguese, Xinhua news agency for Chinese,[3] to mention just a few.

Independent media (news sources free of influence by government or corporate interests), have grown during the last two decades alongside the development of the Internet and reach nowadays a considerable wider audience, thus becoming more influential than they were previously. However, to the best of our knowledge, there are no corpora available from independent media sources.

This paper describes the development of a collection of parallel and monolingual news corpora from independent media. We envisage the resulting novel corpora to be useful in, at least, the following two scenarios:

- Machine translation and computer-assisted translation of independent news.
- Corpus studies of independent news, e.g. characteristics of independent media text, differences in language usage in news between corporate, government and independent sources, etcetera.

The corpora developed in this work are extracted from the Tlaxcala website,[4] self-defined as the international network of translators for linguistic diversity. This website publishes newstories and their translations in fifteen languages, ranging from highly-resourced (e.g. English) to poorly-resourced (e.g. Tamazight).

## 2. Background

Work on corpora acquisition from the Internet can be divided into two broad categories:

- Generic approaches to crawl text from an a priori unknown set of web domains.
- Ad-hoc crawlers to gather text from an a priori known set of web domains.

The pros and cons of each of these approaches are straightforward: the first one is scalable, and thus can be applied, in theory, to any collection of web domains, while the second one, being tailored to specific web domains, should be able to extract better content (i.e. in terms of precision and/or recall) from those web domains.

Generic approaches to obtain monolingual corpora start the process by sending queries to a search engine in order to obtain seed URLs (Baroni et al., 2009) or by traversing a top-domain (e.g. *.es) and performing language identification to keep the web pages in the language of interest, e.g. Catalan (Boleda et al., 2006).

Once that monolingual corpora have been obtained for a set of languages, one can identify document pairs in order to derive parallel corpora. Systems to perform this process include STRAND (Resnik and Smith, 2003), Bitextor (Esplà-Gomis and Forcada, 2010) and ILSP-FC (Papavassiliou et al., 2013).

Ad-hoc approaches allow the developer to tailor the corpora acquisition process to the specific characteristics of the web domains that are to be targeted. This is viable for domains where one could gather valuable data which would not be possible to be acquired by relying on generic approaches.

[1] http://catalog.ldc.upenn.edu/LDC2000T43

[2] http://catalog.ldc.upenn.edu/LDC95T11

[3] http://catalog.ldc.upenn.edu/LDC2003T09

[4] http://www.tlaxcala-int.org/

Two examples of ad-hoc approaches follow. The SETimes corpus (Tyers and Alperen, 2010) contains parallel corpora for nine languages gathered from newstories found at setimes.com. The OpenSubtitles corpus (Tiedemann, 2009) contains parallel corpora consisting of subtitles. In this case the acquisition process is tailored to the nature of the text, e.g. by using the timing information as part of the alignment algorithm.

## 3. Methodology

We have developed a set of scripts to carry out the different phases of the corpus acquisition from the Tlaxcala website. Each of them takes care of a specific task of the process. The following subsections cover in detail each of these tasks.

### 3.1. Crawling

The first step consists on downloading the relevant contents from the web domain: articles and lists of articles. The articles hold the content we are interested in while the lists of articles specify which articles are translations of each other, as well as specifying the language of each article (see Figure 1). A wrapper around the command-line tool `wget` carries out this task.

### 3.2. Data Conversion

The content downloaded is in HTML format, and thus it is necessary to convert it to plain text. We identify manually the boiler-plate sections of article pages. These are removed, and the remaining part of the page (article content) is converted to plain text with the aid of two python scripts: `html2text.py`[5] performs the conversion while `decode_entities.py` takes care of HTML entities.

### 3.3. Identification of Document Pairs

Document pairs are extracted from the lists of articles (see Figure 1) and subsequently stored in a tabbed text file with four fields: identifier of the original article, language code of the original article, identifier of the translated article and language code of the translated article. In the following sample of the document pair database the original article with identifier 10017 written in French has translations in Spanish (identifier 10018) and Portuguese (identifier 10019).

```
10017 fr 10018 es
10017 fr 10019 pt
```

### 3.4. Corpora Building

The structure devised for document pairs introduced in the previous section allows us to build different corpora according to the user needs. E.g. we might build an English Spanish parallel corpus made only of English–Spanish document pairs where the original article is in English.

In our current work we build (i) monolingual corpora for each of the languages and (ii) parallel corpora for each pair of languages. All the corpora are provided sentence split.

#### 3.4.1. Sentence Splitting

The documents gathered for each language are sentence split with Ulysses.[6] This is a recently developed sentence splitter based on unsupervised learning. For each language we train a splitting model with all the data gathered for that language. Subsequently, we perform sentence splitting on the documents in that language by using Ulysses trained on the splitting model.

#### 3.4.2. Alignment

For producing parallel corpora we carry out one additional step: sentence alignment. We use Hunalign (Varga et al., 2005). For each language pair we gather the corresponding document pairs (see Section 3.3.), concatenate them (as we know the document boundaries we mark them with Hunalign's special character `<p>`) and provide them to Hunalign as its input. Alignment is performed in two phases (Hunalign's `realign` parameter) and we keep only one to one sentence pairs.

For language pairs with more than 25,000 sentence pairs we use Hunalign's partial align functionality. Chunks of up to 25,000 sentence pairs are identified before alignment according to the detection of long chains of correspondences[7] and alignment is then performed independently on each of these chunks. This procedure is a workaround due to Hunalign's growth of required memory with input size (Toral et al., 2012).

## 4. Corpora Statistics

This section presents relevant statistics of the monolingual and parallel corpora that have been built. We have built corpora for the following 15 languages: English (en), Spanish (es), French (fr), German (de), Italian (it), Portuguese (pt), Farsi (fa), Arabic (ar), Greek (el), Turkish (tr), Swedish (sv), Tamazight (ber), Catalan (ca), Russian (ru) and Esperanto (eo).

Table 1 shows the number of sentences, tokens, number of types and type-token ratio for the 15 monolingual corpora. The size of the monolingual corpora range from 1,572 sentences and 30,516 tokens (Esperanto) to 203,040 sentences and almost 5 million tokens (English). In terms of type-token ratio, the values range from .0234 (English) to .2771 (Esperanto).

Table 2 shows the number of articles per language, both as originals and as translations, and it also shows which percentage of the articles collected from the website are in each language. It is interesting to note, for example, that two thirds of the English articles are originals and just the remaining third are translations into this language. Conversely, for all the remaining languages except for Arabic and Russian, there are more translations than originals. All in all there are almost five thousand articles as originals and above six thousand as translations.

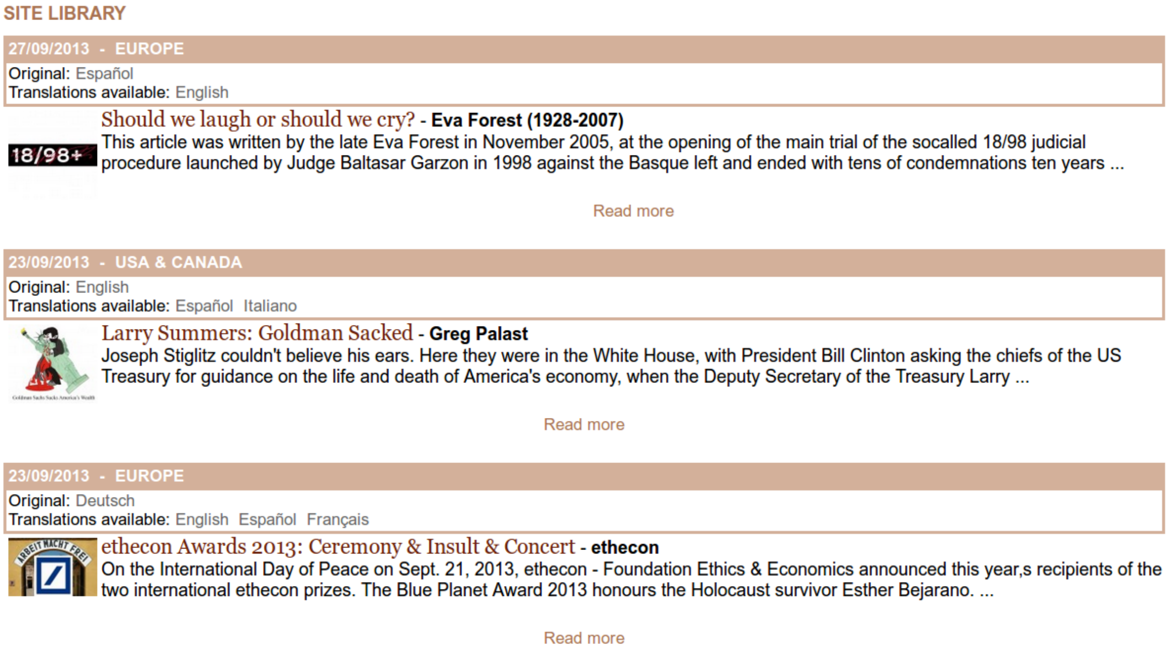
Figure 2 provides a graphic representation of the percentages of articles per language (last column of Table 2).

Finally, we show quantitative information regarding the parallel corpora produced per language pair in Table 3. The

[5] https://github.com/aaronsw/html2text

[6] https://github.com/sortiz/ulysses-sentence-splitter/releases/tag/v0.1

[7] http://mokk.bme.hu/resources/hunalign/



SITE LIBRARY

27/09/2013 - EUROPE
Original: Español
Translations available: English

Should we laugh or should we cry? - **Eva Forest (1928-2007)**
This article was written by the late Eva Forest in November 2005, at the opening of the main trial of the socalled 18/98 judicial procedure launched by Judge Baltasar Garzon in 1998 against the Basque left and ended with tens of condemnations ten years ...

Read more

23/09/2013 - USA & CANADA
Original: English
Translations available: Español Italiano

Larry Summers: Goldman Sacked - **Greg Palast**
Joseph Stiglitz couldn't believe his ears. Here they were in the White House, with President Bill Clinton asking the chiefs of the US Treasury for guidance on the life and death of America's economy, when the Deputy Secretary of the Treasury Larry ...

Read more

23/09/2013 - EUROPE
Original: Deutsch
Translations available: English Español Français

ethecon Awards 2013: Ceremony & Insult & Concert - **ethecon**
On the International Day of Peace on Sept. 21, 2013, ethecon - Foundation Ethics & Economics announced this year,s recipients of the two international ethecon prizes. The Blue Planet Award 2013 honours the Holocaust survivor Esther Bejarano. ...

Read more

Figure 1: Snapshot from the list of articles. For each article, there are links to the original and its translation(s).

| Lang | # sentences | # tokens | # types | Ratio |
|---|---|---|---|---|
| ar | 9,507 | 245,881 | 48,213 | .1960 |
| ber | 4,029 | 122,008 | 15,339 | .1257 |
| ca | 2,216 | 58,242 | 9,287 | .1594 |
| de | 91,519 | 1,911,770 | 120,471 | .0630 |
| el | 9,117 | 225,228 | 29,190 | .1296 |
| en | 203,040 | 4,904,716 | 114,894 | .0234 |
| eo | 1,572 | 30,516 | 8,456 | .2771 |
| es | 175,157 | 4,677,109 | 141,958 | .0303 |
| fa | 16,089 | 412,766 | 27,005 | .0654 |
| fr | 128,782 | 3,566,861 | 107,422 | .0301 |
| it | 63,385 | 1,765,234 | 75,818 | .0429 |
| pt | 56,348 | 1,459,815 | 72,840 | .0498 |
| ru | 4,583 | 85,975 | 20,165 | .2345 |
| sv | 6,419 | 116,158 | 20,916 | .1800 |
| tr | 6,271 | 103,613 | 26,108 | .2519 |

Table 1: Number of sentences, tokens, types and type-token ratio in the monolingual corpora for each language.

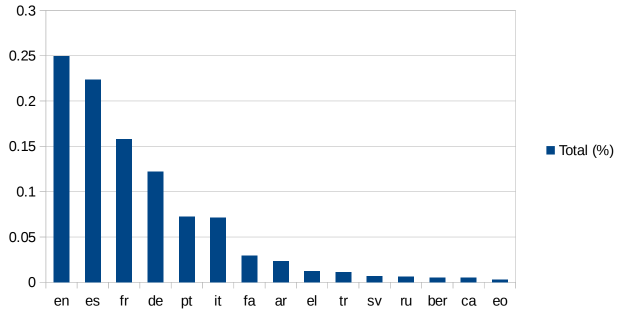


Figure 2: Percentage of articles in the corpus by language.

| Lang | # original | # translation | Total | Total (%) |
|---|---|---|---|---|
| en | 2,047 | 732 | 2,779 | 24.97% |
| es | 1,135 | 1,357 | 2,492 | 22.39% |
| fr | 507 | 1,250 | 1,757 | 15.78% |
| de | 377 | 981 | 1,358 | 12.20% |
| pt | 86 | 720 | 806 | 7.24% |
| it | 179 | 614 | 793 | 7.12% |
| fa | 88 | 239 | 327 | 2.94% |
| ar | 158 | 101 | 259 | 2.33% |
| el | 38 | 100 | 138 | 1.24% |
| tr | 15 | 114 | 129 | 1.16% |
| sv | 22 | 52 | 74 | 0.66% |
| ru | 45 | 26 | 71 | 0.64% |
| ber | 3 | 56 | 59 | 0.53% |
| ca | 17 | 40 | 57 | 0.51% |
| eo | 8 | 24 | 32 | 0.29% |
| Total | 4,725 | 6,406 | 11,131 | 100.00% |

Table 2: Number of documents per language as original, translation, overall and the percentage of articles in the corpus in that language.

upper diagonal shows the number of sentence pairs per language pair. The lower diagonal shows the number of tokens per language pair (this figure corresponds to summing up the number of tokens in the document pairs for both languages).

The largest parallel corpus is that for English–Spanish, containing 66,837 sentence pairs and 3,344,998 tokens. Another language pair, Spanish–French, contains more than 2 million tokens. There are several other language pairs with more than 1 million tokens: English–German, English–French, English–Italian and English–Portuguese.

| - | ar | ber | ca | de | el | en | eo | es | fa | fr | it | pt | ru | sv | tr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ar | - | 73 | 71 | 722 | 80 | 2637 | 209 | 1,759 | 235 | 2,151 | 723 | 345 | 2 | 55 | 45 |
| ber | 3,918 | - | 0 | 202 | 0 | 1697 | 83 | 311 | 33 | 733 | 61 | 76 | 0 | 0 | 0 |
| ca | 1,532 | 0 | - | 74 | 18 | 388 | 0 | 1,646 | 0 | 294 | 71 | 37 | 0 | 0 | 31 |
| de | 26,564 | 7,466 | 4,316 | - | 666 | 41,380 | 30 | 9,784 | 6,896 | 16,749 | 1,556 | 460 | 599 | 1169 | 145 |
| el | 1,714 | 0 | 881 | 27,101 | - | 2,721 | 11 | 2,107 | 48 | 2,259 | 778 | 294 | 0 | 0 | 31 |
| en | 113,621 | 100,439 | 19,317 | 1,809,821 | 119,084 | - | 420 | 66,837 | 3136 | 34,640 | 28,560 | 29,662 | 1,209 | 2,782 | 2,155 |
| eo | 2,474 | 4,558 | 0 | 792 | 347 | 24,667 | - | 559 | 8 | 385 | 20 | 0 | 8 | 8 | 0 |
| es | 109,881 | 21,399 | 85,579 | 438,971 | 94,774 | 3,344,998 | 15,657 | - | 411 | 36,951 | 10,621 | 6,163 | 2,028 | 232 | 1,494 |
| fa | 11,456 | 1,555 | 0 | 272,552 | 2,853 | 147,753 | 375 | 20,130 | - | 1,396 | 0 | 2 | 29 | 10 | 0 |
| fr | 110,866 | 39,063 | 13,183 | 784,100 | 126,708 | 1,727,174 | 20,180 | 2,102,866 | 49,882 | - | 11,882 | 5,690 | 451 | 1,058 | 331 |
| it | 33,790 | 3,990 | 4,772 | 70,914 | 38,414 | 1,550,098 | 1,075 | 604,052 | 0 | 663,225 | - | 866 | 318 | 3 | 285 |
| pt | 14,027 | 4,704 | 2,413 | 18,301 | 10,720 | 1,520,698 | 0 | 316,966 | 21 | 315,027 | 50,908 | - | 525 | 2 | 2 |
| ru | 19 | 0 | 0 | 22,975 | 0 | 50,361 | 348 | 83,871 | 319 | 16,706 | 14,677 | 21,320 | - | 0 | 0 |
| sv | 612 | 0 | 0 | 48,416 | 0 | 96,207 | 329 | 8,860 | 311 | 44,157 | 129 | 16 | 0 | - | 10 |
| tr | 587 | 0 | 1,718 | 4,002 | 335 | 65,590 | 0 | 59,609 | 0 | 14,998 | 13,281 | 18 | 0 | 243 | - |

Table 3: Number of sentence pairs and tokens per language pair.

## 5. Conclusions

The paper has presented corpora from the domain of independent news acquired from the Tlaxcala website. We have built monolingual corpora for 15 languages and parallel corpora for all the combinations of those 15 languages.

The paper has detailed the acquisition process and it also has presented detailed statistics of the produced corpora, concerning mainly quantitative dimensions such as the size of corpora per language (for the monolingual corpora) and per language pair (for the parallel corpora).

To the best of our knowledge, these are the first publicly-available parallel and monolingual corpora for the domain of independent news. Moreover, we have built parallel corpora for languages for which only very limited such resources exist (e.g. Tamazight).

All the content in Tlaxcala is publicly available,[8] and so it is the corpora[9] presented in this paper.

As a side effect of building the corpora, we have created sentence splitting models for Ulysses, an unsupervised sentence aligner. These models are made publicly available together with the corpora. This is deemed to be a relevant further contribution, as hitherto there are no splitters available for some of the languages tackled in this work.

## 6. Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement PIAP-GA-2012-324414 (Abu-MaTran).

## 7. References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Boleda, G., Bott, S., Castillo, C., Meza, R., Badia, T., and Lpez, V. (2006). Cucweb: a catalan corpus built from the web. In Kilgarriff, A. and Baroni, M., editors, *2nd Web as Corpus Workshop at EACL'06*, April.

Esplà-Gomis, M. and Forcada, M. (2010). Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93:77–86.

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, August. Association for Computational Linguistics.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380, September.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Toral, A., Poch, M., Pecina, P., and Thurmair, G. (2012). Efficiency-based evaluation of aligners for industrial applications. In *EAMT 2012: Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 57–60, Trento, Italy.

Tyers, F. M. and Alperen, M. S. (2010). SETimes: A parallel corpus of balkan languages. In *Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages at the Language Resources and Evaluation Conference*, pages 1–5.

Varga, D., Nmeth, L., Halcsy, P., Kornai, A., Trn, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Proceedings of RANLP*, pages 590–596, Borovets, Bulgaria.

[8] http://www.tlaxcala-int.org/copyleft.asp

[9] http://www.computing.dcu.ie/~atoral/resources/tlaxcala