*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3932–3936
Marseille, 11–16 May 2020

# Building the Spanish-Croatian Parallel Corpus

**Bojana Mikelenić, Marko Tadić**
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb
bmikelen@ffzg.hr, marko.tadic@ffzg.hr

### Abstract

This paper describes the building of the first Spanish-Croatian unidirectional parallel corpus, which has been constructed at the Faculty of Humanities and Social Sciences of the University of Zagreb. The corpus is comprised of eleven Spanish novels and their translations to Croatian done by six different professional translators. All the texts were published between 1999 and 2012. The corpus has more than 2 Mw, with approximately 1 Mw for each language. It was automatically sentence segmented and aligned, as well as manually post-corrected, and contains 71,778 translation units. In order to protect the copyright and to make the corpus available under permissive CC-BY licence, the aligned translation units are shuffled. This limits the usability of the corpus for research of language units at sentence and lower language levels only. There are two versions of the corpus in TMX format that will be available for download through META-SHARE and CLARIN ERIC infrastructure. The former contains plain TMX, while the latter is lemmatised and POS-tagged and stored in the aTMX format.

**Keywords:** written corpus, parallel corpus, Spanish, Croatian

## 1. Introduction

A parallel or translation corpus can be bilingual or multilingual (texts in one language and their translation to one or more languages) and unidirectional (source language→target language), bidirectional (language 1↔language 2) or multidirectional (source language→multiple target languages) (McEnery, Xiao and Tono, 2006, 48). This paper describes the process of building a bilingual unidirectional (Spanish→Croatian) parallel corpus, where a well-resourced language (Spanish) is paired with an under-resourced language (Croatian). The possible uses of parallel corpora in linguistic and NLP research are well known and span from contrastive and translation analyses to bilingual lexica extraction, and they could also be used for training statistical or neural machine translation models.

The corpus we are describing here was built in collaboration between two departments of the Faculty of Humanities and Social Sciences of the University of Zagreb, namely Department of Romance Languages and Literatures and Department of Linguistics.

The paper is organised as follows: Section 2 summarizes previous work on parallel corpora that includes this language pair; in Section 3 we give a brief overview of the corpus parameters; Section 4 is dedicated to the corpus composition, including sampling procedure (4.1) and output format (4.2); Section 5 describes the corpus processing and Section 6 gives a conclusion and indicates possible future research directions.

## 2. Previous work

To the best of our knowledge, this is the first direct Spanish-Croatian parallel corpus collected as such and the only previous work we can make a reference to are larger multilingual text collections, which allow for extraction of parallel bitexts in these two languages:

- OpenSubs (Tiedemann, 2012), issues 2016[1] and 2018[2], the latter being a new cleaner version of the collection, with better language checking and enhancements in alignment (Lison and Tiedemann, 2016);
- European Union large-scale multilingual language technology resources described in Steinberger et al. (2014)[3], most importantly JRC-Acquis (Steinberger et al., 2006)[4].
- DGT TMs[5] are regularly published on a yearly basis and they consist of translations from different EU-bodies since 2007, with Croatian appearing only from 2014.
- Tilde MODEL – ECB[6] is one of the corpora that form a part of the Tilde MODEL Corpus – Multilingual Open Data for European Languages (Rozis and Skadins, 2017). It is compiled from the European Central Bank web site, which is multilingual.

These collections are valuable resources for different kinds of studies and tasks, but we have to bear in mind that they were created and processed automatically and because of that they can in different quantities contain noise and errors (spelling errors, errors in sentence alignment, missing or wrong diacritics, wrong ISO-code attachment, etc. were found in e.g. OpenSubs), even though efforts are constantly being made for their improvement. The specific nature of the texts that are in these collections also needs to be considered when using these resources. In regards to the type of language, while texts in OpenSubs concern a large variety of topics, EU corpora pertain to a specific legislative style and EU-legal

[1] http://opus.nlpl.eu/OpenSubtitles2016.php

[2] http://opus.nlpl.eu/OpenSubtitles2018.php

[3] https://ec.europa.eu/jrc/sites/jrcsh/files/2014_08_LRE-Journal_JRC-Linguistic-Resources_Manuscript.pdf

[4] Part of Croatian translation aligned with the equivalent English documents is available through META-SHARE: http://meta-share.ffzg.hr/repository/browse/croatian-translations-ofacquis/547866326c1811e28a985ef2e4e6c59e6758e8d15e7a445e9471e185a758b50c/.

[5] See at: https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory#More%20details%20/%20Reference%20publication.

[6] https://tilde-model.s3-eu-west-1.amazonaws.com/Tilde_MODEL_Corpus.html

terminology, which can distort results obtained from using such data. Both are also highly conditioned by their format of appearance: a dialog in a movie and the allowed subtitle length to translate it, specific sentence structure and phrasing in a legislative text, etc. Furthermore, while texts in the EU corpora were produced by professional translators, the ones in OpenSubs were created by unknown individuals with unknown language and translation skills. In regards to the OpenSubs corpora, there can also be a difference in the variety of Spanish (Spain vs. Latin America, or even depending on a country). Lastly, a specific problem with these extracted Spanish-Croatian bitexts is that one is not translated from the other, or in other words, texts in both languages are usually translations of source texts in another language (e.g., source language in EU corpora tends to be English or German).

Our corpus is different because we worked with original Spanish literary texts and published Croatian translations, done by professional translators. In this way, there is a better chance of high quality translation, but since literature uses highly idiosyncratic language, conclusions drawn from research on this data do not necessarily apply for other text types (Lawson, 2001, 294). We have tried to mitigate this by using novels instead of short stories, essays or poetry. While the sizes of parallel corpora for under-resourced languages that involve English as one of languages in a pair can sometimes be measured in millions of TUs, for other language pairs it is not always so. In this respect, the size of this direct Spanish-Croatian parallel corpus, regardless of the limitation of the text type, can still be quite useful for a number of applications and research tasks.

The collection process was manual, whereas conversion and alignment were done automatically and then manually post-checked and corrected. Some of the texts were received in the PDF format and needed to be converted to plain text, so that is why this additional manual correction was completed before alignment. In this way, we believe we have compiled a more reliable resource for this language pair that is also as noise-free as possible.

In addition to being a new language resource, in itself especially valuable for Croatian language technology, we would like to emphasise that this corpus will allow for new research to be conducted about Croatian language, which until now was mostly paired with English or other Slavic languages (e.g., Bulgarian[7], Czech[8] or Macedonian[9]).

## 3. Corpus parameters

This is a unidirectional (spa→cro) parallel corpus, composed of synchronic fictional prose texts in Spanish and their translation to Croatian. The corpus has a total of 2,092,707 tokens (see Table 1 for statistics), with approximately 1 Mw for each language, although the Spanish part amounts to 52,74% and the Croatian part to 47,25% of the corpus. This text contraction of 11,61% after translation is due to more synthetic expression of morphosyntactic features in Croatian as opposed to Spanish, namely, lack of obligatory article or other determinant occurring with nouns, with cases expressed inflectionally with different endings and therefore lighter preposition usage.

| | **Spanish** | **Croatian** |
|---|---|---|
| Tokens | 1,103,789 | 988,918 |
| Sentences (TUs) | 71,778 | |
| Samples | 11 | |
| Average TUs/sample | 6,525.27 | |

Table 1. Statistics of Spanish-Croatian Parallel Corpus

## 4. Corpus composition

### 4.1 Sampling procedure

Having in mind that Spanish is one of the world’s most spoken languages, finding source texts translated and published even to a small language such as Croatian was not too difficult a task. Nevertheless, during the sampling procedure, we followed a number of criteria: availability of text in digital form, language as neutral/general as possible, contemporary texts written by Spanish authors (not Latin American), Croatian translations done by as many different translators as possible.

We were fortunate that *Fraktura*, a publishing house that publishes a large part of Croatian translations of Spanish literary works, allowed us to use the translations for research purposes. As was previously stated, we opted for novels rather than short forms, poetry or essays to lessen the features typical for those shorter types of texts (idiosyncratic language, poetic expressions, very specific terminology, (in-)formality, etc.), even though novels of course still bear certain characteristics of the style, which needs to be taken into account when using the data. Because Spanish is spoken in numerus countries around the world, it has many varieties and dialects. Roughly, we can distinguish between Spanish spoken in Spain (European or Peninsular Spanish) and Spanish spoken in Latin or Hispanic America (or Spanish of the Americas). They especially differ in pronunciation and vocabulary, but there are also certain differences in grammar (e.g. use of prepositions, second-person personal pronouns and verb forms). This is why we have decided that, for now, we will include in our corpus only texts written by Spanish authors. Lastly, it was important to include texts translated by as many translators as possible, to mitigate some specific or personal language use by translators.

The corpus contains eleven Spanish novels in total, all published between 1999 and 2008 and their Croatian translations by six different translators published between 2006 and 2012. The novels, included in the corpus in their entirety, are as follows:

- Spanish texts: Barceló, E. (2003). *El secreto del orfebre*. Ediciones Lengua de Trapo; Mendoza, E. (2008). *El asombroso viaje de Pomponio Flato*. Barcelona: Seix Barral; Ruiz Zafón, C. (2001). *La sombra del viento*. Barcelona: Teide; Martín, E. & Carranza, A. (2007). *La Clave Gaudí*. Barcelona: Plaza y Janes; Ruiz Zafón, C. (2008) *El Juego del Ángel*. Barcelona: Editorial Planeta; Ruiz Zafón, C. (1999). *Marina*. Barcelona: Edebé; Cercas, J. (2001). *Soldados de Salamina*. Barcelona: Tusquets Editores;

[7] Koeva et al. (2012).

[8] InterCorp v11: https://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze11

[9] Cebović and Tadić (2016).

Garrido, A. (2008). *La Escriba*. Barcelona/Madrid: Ediciones B; Calderón, E. (2006). *El mapa del creador*. Barcelona: Editorial Roca; Palma, F.J. (2008). *El mapa del tiempo*. Sevilla: Algaida; Fortes, S. (2009). *Esperando a Robert Capa*. Barcelona: Editorial Planeta.

- Croatian texts: Barceló, E. (2007). *Zlatarova tajna*. Zagreb: Fraktura; Mendoza, E. (2011). *Čudesno putovanje Pomponija Flata*. Zagreb: Fraktura; Ruiz Zafón, C. (2006). *Sjena vjetra*. Zagreb: Fraktura; Martín, E. & Carranza, A. (2009). *Gaudijev ključ*. Zagreb: Fraktura; Ruiz Zafón, C. (2009). *Anđelova igra*. Zagreb: Fraktura; Ruiz Zafón, C. (2011). *Marina*. Zagreb: Fraktura; Cercas, J. (2008). *Salaminski vojnici*. Zagreb: Fraktura; Garrido, A. (2011). *Pisarica*. Zagreb: Fraktura; Calderón, E. (2008). *Stvoriteljeva karta*. Zagreb: Fraktura; Palma, F.J. (2012). *Vremenska karta*. Zagreb: Fraktura; Fortes, S. (2012). *Čekajući Roberta Capu*. Zagreb: Fraktura.

### 4.2 Output format

The Croatian publisher agreed to the usage of these texts for research purposes, but since we did not receive response from all of the Spanish publishers and authors contacted and made aware of the project, we have decided to limit the availability of language units to the level of a sentence. The aligned TUs were therefore shuffled, so the reconstruction of the original order of sentences in either language side is not possible. In this way, no research can be conducted above the sentence level, but the copyrights remain protected. The corpus will be downloadable through META-SHARE[10] and CLARIN ERIC[11] with CC-BY licence. Two versions of the corpus will be available for download in TMX format: the plain version, without any annotation, and the enriched version.

## 5. Corpus processing

All Croatian texts were received in watermarked PDF format and converted with a commercial ABBYY FineReader[12] into DOCX for easier manual correction. The correction included removing pictures (watermarks on empty pages), deleting generated text boxes (e.g. page and chapter numbers), correcting spelling errors and missing diacritics, etc. We then proceeded to convert these documents to simple UTF8 encoded text files, saving them as such directly from Microsoft Word. For Spanish texts received in PDF format, the same procedure was applied, whereas those obtained in EPUB were converted into TXT by a simple Python program which uses the EbookLib Python library[13]. In the resulting TXT files for both languages only plain body text was kept, so the documents were stripped of titles, subtitles, chapter numbers, footnotes, text emphasis, etc. If there were any parts that appeared in a specific edition and were not translated to Croatian or were added only in the Croatian edition (preface, reviews, information about the author, etc.), those were removed as well.

The TXT files were then segmented and aligned on the sentence level using LF-Aligner[14], an open source tool that relies on Hunalign[15] (Varga et al., 2005) for automatic sentence pairing. LF-Aligner generates an XLSX file that allows for manual post-check and correction, after which it creates both TXT and TMX files. The sentence segmentation and alignment was manually checked, but the program performed quite well, so there were not a lot of errors. We didn't track or count the corrections since our primary aim was not to evaluate the quality of the alignment tool. After this step the clean TMX version of the corpus was produced.

```
</tu>
<tu tuid="36">
        <tuv xml:lang="spa">
                <seg>Fui atravesando velo tras
               velo hasta llegar al final del
               corredor, donde se abría una gran
               sala en penumbra.</seg>
        </tuv>
        <tuv xml:lang="hrv">
                <seg>Prolazio sam veo za velom dok
               nisam došao do kraja hodnika gdje
               se otvarala velika dvorana u
               polumraku.</seg>
        </tuv>
</tu>
<tu tuid="37">
        <tuv xml:lang="spa">
                <seg>Se estará preguntando por qué
               lo hice.</seg>
        </tuv>
        <tuv xml:lang="hrv">
                <seg>Sigurno se pitate zašto sam
               to učinio.</seg>
        </tuv>
</tu>
<tu tuid="38">
        <tuv xml:lang="spa">
                <seg>Y qué niña, oiga, para cortar
               el tráfico.</seg>
        </tuv>
        <tuv xml:lang="hrv">
                <seg>A kakva li je samo cura! Da
               ti pamet stane.</seg>
        </tuv>
</tu>
<tu tuid="39">
        <tuv xml:lang="spa">
                <seg>Wells se acercó a ella y
               admiró lleno de asombro aquel
               trabajo exquisito.</seg>
        </tuv>
        <tuv xml:lang="hrv">
                <seg>Wells joj je prišao i
               začuđeno stao promatrati taj
               izuzetno precizni rad.</seg>
        </tuv>
</tu>
```

Figure 1: Example from the Spanish-Croatian Parallel Corpus in TMX format

In the following step we applied the freely available tool Freeling[16] (Padró, 2011) for POS-tagging and lemmatisation of the Spanish part and Croatian Pipeline[17] developed during the project CESAR[18] for POS-tagging and lemmatisation of the Croatian part.

[10] http://www.meta-share.org/

[11] https://www.clarin.eu/

[12] https://www.abbyy.com/en-eu/finereader/

[13] https://pypi.org/project/EbookLib/

[14] https://sourceforge.net/projects/aligner/

[15] https://github.com/danielvarga/hunalign

[16] http://nlp.lsi.upc.edu/freeling

[17] http://lt.ffzg.hr:9090/demo/

[18] http://project-cesar.net. See also the Croatian Language Web Services (hrWS) at META-SHARE.

```
<tu tuid="51266">
<tuv xml:lang="spa_tag">
<seg><![CDATA[
     <s>
        1 Había       haber     VMII3S0
        2 algo        algo      PI0CS00
        3 doloroso    doloroso  AQ0MS00
        4 en          en        SP
        5 su          su        DP3CSN
        6 expresión   expresión NCFS000
        7 ,           ,         Fc
        8 como        como      CS
        9 si          si        CS
       10 no          no        RN
       11 consiguiera conseguir VMSI3S0
       12 dejar       dejar     VMN0000
       13 me          me        PP1CS00
       14 marchar     marchar   VMN0000
       15 ,           ,         Fc
       16 como        como      CS
       17 si          si        CS
       18 ella        él        PP3FS00
       19 sintiera    sentir    VMSI3S0
       20 también     también   RG
       21 la          el        DA0FS0
       22 mordedura   mordedura NCFS000
       23 de          de        SP
       24 el          el        DA0MS0
       25 arpón       arpón     NCMS000
       26 que         que       PR0CN00
       27 me          me        PP1CS00
       28 mataba      matar     VMII3S0
       29 .           .         Fp
     </s> ]]>
</seg>
</tuv>
<tuv xml:lang="hrv_tag">
<seg><![CDATA[
     <s>
     1    Bilo     biti       V   0    Pred
     2    je       biti       V   1    AuxV
     3    nečeg    nešto      P   1    Sb
     4    bolnog   bolan      A   1    Pnom
     5    u        u          S   1    AuxP
     6    njenu    njen       P   7    Atr
     7    izrazu   izraz      N   5    Adv
     8    ,        ,          Z   10   AuxX
     9    kao      kao        C   10   AuxY
     10   da       da         C   1    Sub_Sb
     11   me       ja         P   13   Obj
     12   ne       ne         Q   13   AuxV
     13   uspijeva uspijevati V   10   Pred_Co
     14   pustiti  pustiti    V   13   Obj
     15   da       da         C   13   Sub_Obj
     16   odem     otići      V   15   Pred
     17   ,        ,          Z   19   AuxX
     18   kao      kao        C   19   AuxY
     19   da       da         C   13   Sub_Obj
     20   i        i          C   18   AuxY
     21   ona      oni        P   22   Sb
     22   osjeća   osjećati   V   20   Pred_Co
     23   ujed     ujed       N   22   Sb
     24   osti     osti       N   23   Atr
     25   koje     koji       P   24   Sub_Atr
     26   su       biti       V   28   AuxV
     27   me       ja         P   28   Obj
     28   ubijale  ubijati    V   25   Pred
     29   .        .          Z   0    AuxK
     </s> ]]>
</seg>
</tuv>
</tu>
```

Figure 2: Example from the enriched Spanish-Croatian Parallel Corpus in aTMX format

The corpus annotation was formatted following the CoNLL-like format of verticalised corpus with additional columns for each new layer of annotation. This format was wrapped in XML elements that form a typical TMX structure and in this way, we produced the version of the corpus in aTMX format (Brito et al. 2014) preserving the alignment at sentence level and providing the additional annotation. Although aTMX format is not accepted as a standard, we opted for this solution because it was converted from a standard TMX in a straightforward way and the CoNLL-like format of annotations is still preserved. Both versions will be available at major language resources repositories (e.g. META-SHARE, CLARIN and probably ELG and ELRC-SHARE as well).

## 6. Conclusion and future work

We have described the composition and processing of the Spanish-Croatian parallel corpus that is available in two versions: in TMX format without any annotation and in aTMX format, enriched with POS-tags, lemmas and in the case of Croatian also dependency parsing information. In the future, we would like to be able to publish this corpus with the original sentence order, to allow for research of language units above the sentence level. In addition, we would like to add more texts into the corpus, since there are more Spanish texts available in Croatian translation. It would also be most useful to add different types of texts and texts in Spanish written by Latin American authors that are translated to Croatian.

Since for both languages, UD language resources exist and the CoNLL-U format is defined, in the next iteration we will try to convert the files into UD-compatible format as well. However, right now the dependency parser used for Croatian is following the analytic-level PDT v2.0 tagset adapted for Croatian and is currently not UD-compatible.

By compiling this language resource we can now start checking how much will the existing MT systems improve by incremental training of the existing translation models. However, most of the trained MT systems for this language pair were trained with SMT methods, while now the NMT methods are considered to provide a higher translation quality. In this respect, we can't straightforwardly compare the grade of improvement before retraining of the baseline models using NMT methods is done.

Furthermore, we plan to allow the search of the corpus, including it in one of the available online platforms (e.g. NoSketch Engine)[19]. This step is especially important to make the corpus accessible to language and linguistics professors, students and other professionals.

## 7. Bibliographical References

Brito, R., Almeida, J. J. and Simões, A. (2014). Processing Annotated TMX Parallel Corpora. In IberSpeech 2014, VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop, pages 188-197, Las Palmas.

Cebović, I. and Tadić, M. (2016). Building the Macedonian-Croatian Parallel Corpus. In Nicoletta Calzolari (Conference Chair), et al. (eds.) Proceedings

[19] https://nlp.fi.muni.cz/trac/noske

of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4241–4244, Portorož, Slovenia, European Language Resource Association (ELRA).

Koeva, S., Stoyanova, I., Dekova, R., Rizov, B., Genov, A. (2012). Bulgarian X-language Parallel Corpus. In: Nicoletta Calzolari (Conference Chair) et al. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 51-62, Istanbul, Turkey, European Language Resources Association (ELRA).

Lawson, A. (2001). Collecting, aligning and analysing parallel corpora. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small Corpus Studies and ELT: theory and practice*. Amsterdam/Philadelphia: John Benjamins, pp. 279–309.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Nicoletta Calzolari (Conference Chair), et al. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 923–929, Portorož, Slovenia, European Language Resource Association (ELRA).

McEnery, T., Xiao, R. and Tono, Y. (2006). Corpus-Based Language Studies. Routledge, London and New York.

Padró, L. (2011). Analizadores Multilingües en FreeLing. *Linguamatica*, *3*(2):13–20.

Rozis, R. and Skadins, R. (2017). Tilde MODEL - Multilingual Open Data for EU Languages. Proceedings of the 21th Nordic Conference of Computational Linguistics NODALIDA 2017, Gothenburg, Sweden.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), pages 2142–2147, Genoa, Italy, European Language Resource Association (ELRA).

Steinberger R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., and Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*. 48(4):679–707.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), et al. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey, European Language Resource Association (ELRA).

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Nemeth, L., and Tron, V. (2005). Parallel corpora for medium density languages. In G. Angelova, K. Bontcheva, R. Mitkov, Nicolas Nicolov, & Nikolai Nikolov (eds.), Proceedings of the RANLP 2005 (Recent Advances in Natural Language Processing), pages 590–596, Borovets, Bulgaria.