# Building the Macedonian-Croatian Parallel Corpus

**Ines Cebović, Marko Tadić**

Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
ines.cebovic@gmail.com, marko.tadic@ffzg.hr

**Abstract**

In this paper we present the newly created parallel corpus of two under-resourced languages, namely, Macedonian-Croatian Parallel Corpus (**mk-hr_pcorp**) that has been collected during 2015 at the Faculty of Humanities and Social Sciences, University of Zagreb. The **mk-hr_pcorp** is a unidirectional (mk→hr) parallel corpus composed of synchronic fictional prose texts received already in digital form with over 500 Kw in each language. The corpus was sentence segmented and provides 39,735 aligned sentences. The alignment was done automatically and then post-corrected manually. The alignments order was shuffled and this enabled the corpus to be available under CC-BY license through META-SHARE. However, this prevents the research in language units over the sentence level.

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

## 1. Introduction

Parallel corpora are needed more and more as the source of bilingual and/or multilingual data for different research tasks in linguistics and Natural Language Processing. The research in contrastive linguistics, bilingual lexicography, translation studies, up to a traning of statistical machine translation models are just few of them. Today in the majority of parallel corpora it is usual that at least one of languages is a well-resourced language, while parallel corpora with two (or more) under-resourced languages are rather rare.

In this paper we present the newly created parallel corpus of two under-resourced languages, namely, Macedonian-Croatian Parallel Corpus (**mk-hr_pcorp**) that has been collected during 2015 in collaboration between two departments of the Faculty of Humanities and Social Sciences, University of Zagreb: Department of South-Slavic Languages and Department of Linguistics.

The paper is composed as follows: in Section 2. we mention the previous work on parallel corpora that include these two languages. In Section 3. we describe corpus parameters of **mk-hr_pcorp** and in Section 4 we provide information about the corpus composition. The Section 5 describes the processing and in Section 6 we give conclusion and mention possible future development.

## 2. Previous work

To the best of our knowledge, there are no parallel texts collected for this language pair in particular, but there are larger multilingual text collections published earlier out of which a parallel Macedonian-Croatian bitexts could be extracted:

- SETimes corpus, published in OPUS (Tiedemann, 2009) [1] and (Tyers & Alperen, 2010), which was later corrected and cleaned up (Ljubešić, 2013) [2] of residual HTML code, wrong language identification and missing diacritics[3];
- OpenSubs (Tiedemann, 2012), issues 2012[4] and 2013[5].

These large multilingual text collections were created by automatical crawling or conversion from digital texts available in different input formats and then processed for noise removal. Although numerous automatic checking procedures were applied, still within these text collections a substantial amount of errors and noise can be found, starting with the wrong language ISO-codes attachment, up to the systemic errors in encoding, i.e. missing diacritics, or inaccurate alignments.

The **mk-hr_pcorp** differs from other existing parallel text collections for this language pair because it was collected manually, converted and aligned automatically, and then manually post-checked in order to achieve full alignment accuracy. The original texts and their human-made translations have been received from the authors and translators already in the digital form, so the process of conversion was simplified and usual OCR-step with post-OCR correction was not necessary. In this respect this corpus can be considered noise-free and more reliable resource for this language pair than the ones presented in the previous work. This can be considered as an additional corpus quality and it may become increasingly important in the case of using such corpora as the training material for the tools that should not propagate errors from the training to the further levels of processing.[6] Also, this corpus covers a different domain,

[1] http://opus.lingfil.uu.se/ SETIMES.php

[2] http://nlp.ffzg.hr/resources/ corpora/setimes/

[3] We would highly recommend the usage of this corrected version of SETimes corpus.

[4] http://opus.lingfil.uu.se/OpenSubtitles2012.php

[5] http://opus.lingfil.uu.se/OpenSubtitles2013.php

[6] The question of the highest possible quality of language resources used for training is a different topic that, we believe, should be thoroughly discussed in a specialised workshop. Unfortunately, what we can often witness nowadays is the

i.e. differs in genre and style as it departs from the previous newspaper and subtitles text by including only fiction works.

## 3. Corpus parameters

This corpus is an unidirectional (Macedonian→Croatian) parallel corpus that has been collected from the literary works of Macedonian authors and their translations into Croatian. The Macedonian part of the corpus has 531,936 tokens, while the Croatian counterpart encompasses 509,455 tokens. The 4.33% more tokens in the Macedonian part is expected since more analytical features are present in the Macedonian as against the Croatian: e.g. introduction of the article, no cases in nouns, heavier preposition usage, etc. The corpus also encompasses 39,735 aligned sentences, i.e., translation units (TUs).

Since the lower borderline year for the samples in the Croatian National Corpus has been set to 1990 (as explained in Tadić, 2009[7]), i.e. only translations published from that year onwards are included in **mk-hr_pcorp**, the similar condition has been applied to the Croatian half in order to keep the texts synchronic as much as possible and thus keep them compatible with the existing tools for processing Croatian. The Macedonian originals, however span from 1945 up to the present day since the appearance of the first Macedonian grammar was selected as the borderline event.

## 4. Corpus composition

The texts included in **mk-hr_pcorp** were the texts which had the Croatian translations existing and available in digital format. The selection process included all prose types: novels, stories and short stories. The poetry was left out because the translations of poetic texts often exhibit the substantial divergence from the original due to the fact that translators have to follow the metric schemes and/or rhyme, and this also dictates the selection of the translation equivalents as well as word order. In such cases the lexical and co-/con-textually appropriate choices don't get selected in the most expected or neutral way, therefore inclusion of poetic texts could produce more noise. This we see as important particularly in the case of this first manually composed and checked Macedonian-Croatian parallel corpus, that might be used as a baseline in future research. In future versions of this corpus we will consider the classification of samples by genre and text type so that the appropriate information can be attached to each TUs.

In order to reach at least 500 Kw in each language, the sampling procedure was not strict, but all available texts in both languages, that fulfilled the predescribed parameters, were used completely. This approach can be seen often in the cases in the collection of parallel corpora of under-resourced languages. In such cases the quantity preceedes the sampling method in order to keep the corpus size as large as possible.

However, even with all the effort put into the collecting as much parallel texts as possible, the final size of this **v1** of **mk-hr_pcorp** is certainly not sufficient for training of serious statistical machine translation models, but it could be more useful as the source of bilingual linguistic data for translation studies, bilingual lexicography and contrastive linguistic research.

| | **Macedonian** | **Croatian** |
|---|---|---|
| Tokens | 531,936 | 509,455 |
| Sentences (TUs) | 39,735 | |
| Samples | 16 | |
| Average TUs/sample | 2483.44 | |

Table 1. Basic statistics of **mk-hr_pcorp**

Authors and translators have agreed to the usage of their texts for research purposes in the case that the available co-text will be limited. In the parallel concordance no more than a single aligned sentence pair is displayed, so the copyrights remain protected since it is very hard to claim the copyright to a single sentence or its translation. Also, for the distribution version of the corpus, the aligned TUs are scrambled, so no original order of sentences within the texts can be reproduced. In this way the copyrights will be additonally protected while the corpus will be fully downloadable and available through META-SHARE[8] platform for language resources distribution with CC-BY license right after its public presentation. The corpus will be available for download in TMX and Moses formats. The list of original texts, that were used as samples, will be available within the metadata in the META-SHARE record for this language resource.

## 5. Corpus processing

The conversion was performed from .docx files into simple UTF8 encoded text files using simple VBscripts. This procedure stripped off all text-formatting information of titles, subtitles, heads of chapters, text emphasis within paragraphs etc., but kept the paragraph boundaries. After manual checking and correcting for the same number of parallel paragraphs in all samples, the sentence alignment was performed. We used freely available tool NOVA Text Aligner[9] which supports the

plethora of automatically crawled (multilingual) language resources that have been produced in larger than ever sizes, but that have never been observed by a human eye and never checked with scrutiny needed for the fundamental language data such as language resources, particularly ones used for later training of different language tools.

[7] The 1990 was selected because that was the year when the fundamental sociolinguistic change happened for Croatian language. Namely, from one of the official languages of multilingual federal state controlled by a communist regime, it became an official language of an independent democratic state. The differences in language (primarily reflected through changes in orthography, vocabulary and frequency of usage of many lexical and syntactic items that were previously banned), was so wast, that it couldn't have been neglected.

[8] http://www.meta-share.eu

[9] http://www.supernova-soft.com/wpsite/products/text-aligner/

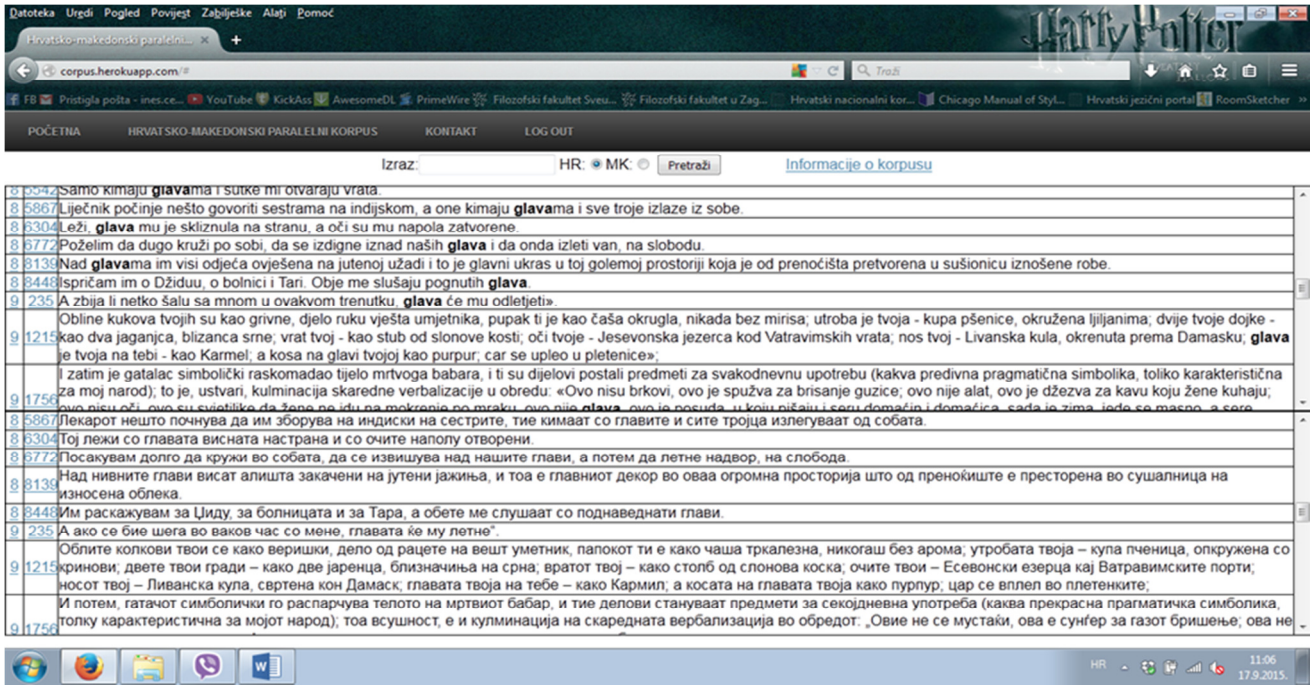


Figure 2: Example of search string "glava" in Croatian. Since the corpus is not yet lemmatised, only the exact string match is possible. The upper half of results displays the Croatian sentences, while the lower half displays their aligned Macedonian counterparts.

UNICODE encoding and performs the automatic sentence segmentation according to the predefined punctuation rules. Since by the Croatian orthographical rules the order numerals written in Arabic numbers obligatory end with a dot, all such cases had to be manually checked to avoid errors in sentence segmentation because ca 24% of these dots are in the same time the fullstops, i.e mark the ordinality and the sentence ending. For Macedonian the segmenter performed remarkably well with its language independent setting. Since the sentence segmentation is a feature of the software package and since we had no information of its method, we didn't perform the thorough evaluation of sentence segmentation process.

```
<tu>
    <tuv xml:lang="mk">
        <seg>Првата квечерина веќе ги
        покриваше планините над Скопје кога,
        сиот задишан, Гордан Коев втрча во
        големата хала на новата железничка
        станица.</seg>
    </tuv>
    <tuv xml:lang="hr">
        <seg>Prvi suton već je prekrivao
        planine nad Skopjem kad je Gordan
        Koev, sav zadihan, utrčao u veliko
        predvorje nove željezničke
        stanice.</seg>
    </tuv>
</tu>
<tu>
    <tuv xml:lang="mk">
        <seg>Станичната зграда од метал и
        затемнето валкано стакло, сега
        изложена на новиот летен бран јулска
        врелина, издишуваше ширејќи мирис на
        амонијак што тешко се источуваше од
        зад цвичавата порта на станичниот
        тоалет.</seg>
    </tuv>
    <tuv xml:lang="hr">
        <seg>Stanična zgrada od metala i
        zatamnjenoga lijevanog stakla,
        izložena novom ljetnom valu
        srpanjske vreline, izdisala je
        širećи miris amonijaka što se
        isparavao kroz škripeća vrata
        staničnoga zahoda.</seg>
    </tuv>
</tu>
<tu>
    <tuv xml:lang="mk">
        <seg>Новата железничка станица беше
        всушност веќе две децении стара
        зграда со приземје, кат и перони на
        уште повисокото ниво.</seg>
    </tuv>
    <tuv xml:lang="hr">
        <seg>Nova je željeznička stanica
        bila, ustvari, već dva desetljeća
        stara zgrada s prizemljem, katom i
        kolosijecima na još višoj
        razini.</seg>
    </tuv>
</tu>
```

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

The selected software provided a simple automatic alignment that had to be manually corrected to ensure fully alignment accuracy. Since the simple aligning method was of of unknown underlying methodology, we didn't evaluate it, but we used it only as a preliminary step in the full alignment process.

After the alignment, the aligned sentence pairs were stored in a database and are currently available for ad hoc on-line search through the Heroku platform[10] using words or phrases as query input. The search is possible in both languages and it results in the list of sentences containing the searched string with their aligned counterparts.

## 6. Conclusions and future work

We have presented a procedure of building a Macedonian-Croatian parallel corpus, its size, time-span, composition and how it was aligned. This is the first version of this corpus and enhanced version is planned.

Although for Macedonian, to the best of our knowledge, there is no automatic POS-tagger available yet, even though the MulTextEast specification for Macedonian has been produced by Katerina Zdravkova (Erjavec, 2010), we would like to process further this parallel corpus with standard tools for POS/MSD-tagging and lemmatisation. For Croatian part we will use the existing language processing Croatian pipeline (Padró et al., 2014) while for Macedonian the similar pipeline yet has to be constructed. We believe that the Macedonian part of this corpus could respresent the valuable resource for the development of these tools as well.

We would also like to test the usage of the proposed aTMX format (Brito et al., 2014) in order to avoid the large XML overhead that other annotated formats introduce. The example with Croatian part already processed can be seen in Figure 3.

```
<tu>
<tuv xml:lang="mk">
<seg><![CDATA[ <s>
  1  Првата
  2  квечерина
  3  веќе
  4  ги
  5  покриваше
  6  планините
  7  над
  8  Скопје
  9  кога
 10  ,
 11  сиот
 12  задишан
 13  ,
 14  Гордан
 15  Коев
 16  втрча
 17  во
 18  големата
 19  хала
 20  на
 21  новата
 22  железничка
 23  станица
 24  .
</s> ]]></seg>
</tuv>
<tuv xml:lang="hr">
<seg><![CDATA[ <s>
  1  Prvi          prvi          A    O      2 Atr
  2  suton         suton         A    O      3 Atr_Co
  3  već           već           C    O      0 Coord
  4  je            biti          V    O      5 AuxV
  5  prekrivao     prekrivati    V    O      3 Pred
  6  planine       planina       N    O      5 Obj
  7  nad           nad           S    O      6 AuxP
  8  Skopjem       Skopje        N    LOC    7 Atr
  9  kad           kad           C    O      8 Sub_Atr
 10  je            biti          V    O      9 Pred
 11  Gordan        Gordan        N    PER   12 Atr
 12  Koev          Koev          N    PER   13 Sb_Ap
 13  ,             ,             Z    O      9 AuxX
 14  sav           sav           A    O     15 Atr
 15  zadihan       zadihan       A    O     13 Pnom_Co
 16  ,             ,             Z    O     13 AuxX
 17  utrčao        utrčati       V    O     16 Pred_Co
 18  u             u             S    O     17 AuxP
 19  veliko        velik         A    O     20 Atr
 20  predvorje     predvorje     N    O     18 Adv
 21  nove          nov           A    O     23 Atr
 22  željezničke   željeznički   A    O     23 Atr
 23  stanice       stanica       N    O     20 Atr
 24  .             .             Z    O      0 AuxK
</s> ]]></seg>
</tuv>
</tu>
```

Figure 3: Example of future POS-tagging and lemmatisation stored in aTMX format. The columns are: 1) token-ID, 2) token, 3) lemma, 4) POS, 5) NE, 6) dependency node, 7) syntactic tag

[10] https://www.heroku.com

## 7. References


Brito, R., Almeida, J. J., Simões, A. 2014. Processing annotated TMX parallel corpora. In: IberSpeech 2014, VIII Jornadas en Tecnologías del Habla and IV Iberian SLTech Workshop, Las Palmas, pp 188-197. [http://ambs.perl-hackers.net/publications/tmxa.pdf]

Erjavec, T. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), pp 2544-2547.

Ljubešić, N. 2013. SETimes – A Parallel Corpus of English and South-East European Languages, [http://nlp.ffzg.hr/resources/corpora/setimes/].

Padró, L. and Agić, Ž. and Carreras, X. and Fortuna, B. and García-Cuesta, E. and Li, Zhixing and Štajner, T. and Tadić, M. 2014. Language Processing Infrastructure in the XLike Project. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014), pp 3811-3816.

Tadić, M. 2009. New version of the Croatian National Corpus. In: Hlaváčková, Dana ; Horák, Aleš ; Osolsobě, Klara ; Rychlý, Pavel (ur.) After Half a Century of Slavonic Natural Language Processing, Masaryk University, Brno, pp 199-205.

Tiedemann, J. 2009. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pp 237-248, John Benjamins, Amsterdam/Philadelphia.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), pp 2214-2218.

Tyers, F. M. and Alperen, M. S. 2010. SETimes: A parallel corpus of Balkan languages. In: Piperidis, S. and Slavcheva, M. and Vertan, C. (eds.) Proceedings of the MultiLR Workshop at the Language Resources and Evaluation Conference, LREC2010, pp 49-53. [http://xixona.dlsi.ua.es/~fran/publications/lrec2010.pdf].