Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) @LREC2022 , pages 50–55
Marseille, June 25, 2022

Challenges of Building Domain-Specific Parallel Corpora from Public Administration Documents

Filip Klubička[1,2], Lorena Kasunić[1], Danijel Blazsetin[1], Petra Bago[1]
[1]University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb, Croatia
Ivana Lučića 3, 10 000 Zagreb, Croatia
[2]Technological University Dublin, ADAPT Centre, Dublin, Ireland
filip.klubicka@adaptcentre.ie, {lorena.kasunic, blazsetin7}@gmail.com, pbago@ffzg.hr

Abstract

PRINCIPLE was a Connecting Europe Facility (CEF)-funded project that focused on the identification, collection and processing of language resources (LRs) for four European under-resourced languages (Croatian, Icelandic, Irish and Norwegian) in order to improve translation quality of eTranslation, an online machine translation (MT) tool provided by the European Commission. The collected LRs were used for the development of neural MT engines in order to verify the quality of the resources. For all four languages, a total of 66 LRs were collected and made available on the ELRC-SHARE repository under various licenses. For Croatian, we have collected and published 20 LRs: 19 parallel corpora and 1 glossary. The majority of data is in the general domain (72 % of translation units), while the rest is in the eJustice (23 %), eHealth (3 %) and eProcurement (2 %) Digital Service Infrastructures (DSI) domains. The majority of the resources were for the Croatian-English language pair. The data was donated by six data contributors from the public as well as private sector. In this paper we present a subset of 13 Croatian LRs developed based on public administration documents, which are all made freely available, as well as challenges associated with the data collection, cleaning and processing.

Keywords: language resources, parallel corpora, machine translation, Connecting Europe Facility, eTranslation, PRINCIPLE

1. Introduction

PRINCIPLE[1] (*Providing Resources in Irish, Norwegian, Croatian and Icelandic for the Purposes of Language Engineering*) was a project funded by the Connecting Europe Facility (CEF) Telecom instrument[2], a project that focused on the identification, collection and processing of language resources (LRs) for four European under-resourced languages (Croatian, Icelandic, Irish and Norwegian) in order to improve translation quality of eTranslation[3], an online machine translation (MT) tool provided by the European Commission (EC). In this paper we present a freely available subset of Croatian LRs developed based on the public administration documents as well as challenges associated with the data collection, cleaning and processing.

The paper is structured as follows: Section 2 discusses the motivation for data collection as well as the objectives of the PRINCIPLE project within which the aforementioned activities took place, while Section 3 presents related work. Section 4 outlines the data collection process of all Croatian LRs with specific attention to the collection of public sector information. Section 5 describes the challenges of the data cleaning and processing we faced. In Section 6 we present statistics for 13 parallel corpora we have collected in three DSI domains (eJustice, eHealth, eProcurement) as well as in the general domain, followed by Section 7 with a conclusion.

2. Motivation and Objectives

In 2015 the EC, with its president at the time Jean-Clause Juncker, announced *A Digital Single Market Strategy for Europe*[4], identifying the Internet and digital technologies as an opportunity to contribute to the economy, create new jobs, and enhance Europe's position as a world leader in the digital economy which will contribute to the European digital transformation. The necessary EU digital transformation has been recognized by the von der Leyen Commission as well making *A Europe fit for the digital age* one of six priorities[5]. We can safely state that this transformation was abruptly accelerated in various social and economic sectors by the COVID-19 pandemic outbreak in 2020.

One way the EC supports the digital transformation and multilingualism is funding of the development of language technologies (resources and tools) of all its official languages as well as additional non-EU languages[6]. However, language technology support differs significantly between languages and language pair combinations. Rehm and Uszkoreit (2012) and Rehm et al. (2014) analyzed the state of language technology for 47 European languages investigating four categories: machine translation, speech processing, text analytics, and speech and text resources. Only the English language has good support in all four categories. All other languages have moderate support, fragmentary support or weak/no support, with the majority of the languages falling into the last category.

PRINCIPLE was a 2-year project (September 2019 - August 2021) funded by the CEF instrument, which focused on the identification, collection and processing of

[1] https://principleproject.eu/

[2] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom

[3] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

[4] The European Commision. Communication from the Commision to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – A Digital Single Market Strategy for Europe. The European Commission : Brussels 2015. 192 final. URL: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52015DC0192&from=EN. (3 Jan 2022)

[5] The European Commision. The European Commission's priorities. https://ec.europa.eu/info/strategy/priorities-2019-2024_en (3 Jan 2022)

[6] For example Icelandic and Norwegian since Iceland and Norway are part of the European Free Trade Association (EFTA), which are part of the EU single market via the European Economic Area (EEA).

LRs for four European under-resourced languages: Croatian, Icelandic, Irish and Norwegian (covering both official varieties Bokmål and Nynorsk) (Way and Gaspari, 2019; Way et al., 2020). The Action was coordinated by Dublin City University, and involved the Faculty of Humanities and Social Sciences of the University of Zagreb, the National Library of Norway, the University of Iceland and Iconic Translation Machines Ltd. The main focus of the Action was on providing high-quality data in order to improve translation quality of eTranslation, an online MT tool provided by the EC, with a specific focus on two DSI domains: eJustice and eProcurement. Due to the COVID-19 pandemic outbreak, the focus was also extended to the eHealth DSI domain during the project duration. In order to verify the quality of the collected LRs, bespoke domain-adapted neural machine translation (NMT) engines were developed. The evaluations of the MT systems built as part of the project show significant improvements in BLEU scores on in-domain test datasets when using the collected project data. In-domain systems using our data outperform the best online systems by as much as 14.7 points (see Table 1). More details on building the NMT systems can be found in Moran et al. (2021). For all four languages, a total of 66 LRs were collected and made available on the ELRC-SHARE repository[7] under various licenses.

Engine	eProcurement	eJustice	eHealth
Iconic Engine	56.3	51.1	52.9
ONLINE1	49.1	37.9	38.2
ONLINE2	45.9	31.7	43.8

Table 1: SBLEU evaluations scores of the various in-domain engines built as part of the PRINCIPLE project.

In this paper we focus on our contribution to Croatian LRs. Based on the aforementioned cross-language comparison (Rehm and Uszkoreit 2012), the Croatian language has weak or no support in three categories: machine translation, speech processing and text analytics, and has fragmentary support for speech and text processing. In a recent overview of the European language technology landscape conducted by Rehm et al. (2020), it is revealed that still no national funding programs exist in Croatia despite Croatian being a technologically underdeveloped language.

As part of the PRINCIPLE Action activities, we have collected and published a total of 20 Croatian LRs, most of which contain the Croatian-English language pair: 19 parallel corpora and 1 glossary, all uploaded to the ELRC-SHARE repository. The majority of the data belongs to the general domain (72 % translation units [TUs]), while the rest belong to the eJustice (23 %), eHealth (3 %) and eProcurement (2 %) DSI domains. The data was donated by seven data contributors from the public and private sector.

In this paper we only provide a brief overview of the full set of Croatian LRs collected in the Action, and focus on 13 freely available Croatian LRs developed from public administration documents, as well as challenges associated with data collection, cleaning and processing.

3. Related Work

PRINCIPLE initially focused on building high-quality parallel corpora within the eJustice and eProcurement DSI domain. However, as a result of the COVID-19 disease pandemic outbreak, in the course of the project implementation, collection of language resources was extended to the eHealth DSI domain as well.

The ELRC-SHARE repository serves as a hub for "documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination and considered useful for feeding the CEF Automated Translation (CEF.AT) platform"[8]. An analysis of the repository reveals that out of 1416 parallel corpora, almost half are pertinent to the eHealth DSI domain (690 i.e. 49%). Out of those 690 parallel corpora, 636 (92%) were uploaded in 2020 or later, while only 54 (8%) were uploaded before the pandemic outbreak. The data has been collected from publicly available portals (e.g. Publications Office of the European Union[9], European Medicines Agency[10], portal[11] of the European Centre for Disease Prevention and Control[12], press corner of the EC, portal of the European Parliament[13], Wikipedia articles on regarding health and COVID-19 domain, etc.) as part of various projects financed by the EC (e.g. various iterations of the European Language Resource Coordination (ELRC)[14] and the Paracrawl[15] project, the EuroPat[16] project, the MaCoCu[17] project, etc.). The aforementioned projects produced the majority of the parallel corpora within the eJustice (396) and the eProcurement (358) DSI domains. In contrast with these previous projects which gathered publicly available data from readily available sources, the data collected as part of the PRINCIPLE project was not publicly available on portals of public administration bodies, but was scattered on their websites (unpaired

[7] https://elrc-share.eu/
On the ELRC-SHARE repository there are two projects related to the PRINCIPLE Action: "PRINCIPLE - Unevaluated" and "PRINCIPLE - Evaluated". Unevaluated data corresponds to LRs that have been created by the project partners from documents donated by data providers. These LRs have been processed and cleaned either by the data providers themselves or by PRINCIPLE project partners, thus ensuring high quality. Evaluated data means that the LRs in question have been used by Iconic Translation Machines Ltd. to develop a range of state-of-the-art bespoke MT engines for early adopters, matching their specific use-cases. It should be noted that these LRs are a subset of the unevaluated LRs and have gone through additional processing, cleaning and evaluation steps.
PRINCIPLE - Unevaluated:
https://elrc-share.eu/repository/search/?q=&selected_facets=projectFilter_exact%3APRINCIPLE%20-%20Unevaluated.
PRINCIPLE - Evaluated:
https://elrc-share.eu/repository/search/?q=&selected_facets=projectFilter_exact%3APRINCIPLE%20-%20Evaluated.

[8] https://elrc-share.eu/
[9] https://op.europa.eu/en/home
[10] https://www.ema.europa.eu/
[11] https://antibiotic.ecdc.europa.eu/
[12] https://www.ecdc.europa.eu/en
[13] https://www.europarl.europa.eu/
[14] https://www.lr-coordination.eu/
[15] https://paracrawl.eu/
[16] https://europat.net/
[17] https://macocu.eu/

parallel documents), while some were not even publicly available.

In other related work, we acknowledge that there have been various endeavors applying diverse methods to collect corpora appropriate for MT engines. Here we present only a small selection of such works related to low-resourced language pairs and/or domains. Váradi et al. (2020) present the Croatian-English Parallel Corpus of Croatian National Legislation consisting of over 1,800 documents developed as part of the MARCELL[18] project. The English translations were exclusively in PDF format, hence the quality of automatic text extraction was diminished. Sentence alignment was performed automatically, and a manual inspection and correction was conducted. The corpus contains 396,984 TUs. Utka et al. (2022) present the English-Lithuanian comparable corpus DVITAS in the cybersecurity domain containing over 1,700 English and 2,500 Lithuaninan documents developed for the automatic bilingual term extraction. The corpus contains 4M words, 2M per language. Ghaddar and Langlais (2020) present a Large Scale French-English Financial Domain Parallel Corpus SEDAR in the low-resourced financial domain based on publicly available documents and information in PDF format. Due to this particular information being strictly forbidden to extract automatically, the authors describe the methodology they have applied for text extraction. The corpus contains 8.6 million high quality sentence pairs.

4. Data Collection

The data used for the development of Croatian LRs was donated by six data contributors from the public sector and one from the private sector:

- the Ministry of Foreign and European Affairs,
- the Central State Office for the Development of the Digital Society,
- the Central State Office for Central Public Procurement,
- the State Commission for Supervision of Public Procurement Procedures,
- the Faculty of Humanities and Social Sciences, University of Zagreb, and
- Ciklopea d.o.o.

The Ministry of Foreign and European Affairs donated documents in various formats (MS Word format, PDF format, TMX format, SDLTM format, HTML format) that were used for the development of five LRs in eJustice and eHealth DSI domains.

The Central State Office for the Development of the Digital Society donated documents in MS Word, PDF and HTML format that were used for the development of five LRs in the eProcurement and eJustice DSI domains, as well as the general domain.

The Central State Office for Central Public Procurement donated documents in MS Word format that were used for the development of two LRs in the eProcurement DSI domain.

The State Commission for Supervision of Public Procurement Procedures donated documents in MS Word format that were used in the development of three LRs in the eProcurement and eJustice DSI domains.

The Faculty of Humanities and Social Sciences University of Zagreb donated four resources. One resource was donated in TMX format in the eJustice DSI domain. The other three resources in the general domain were developed prior to the PRINCIPLE project and are in TXT format[19]. Those resources required acquiring permissions from their developers to use the data for improving the eTranslation system. All four resources from the Faculty of Humanities and Social Sciences did not require any additional processing, and are excluded from further LR descriptions.

Ciklopea d.o.o., a translation and localization company, cleaned, processed and/or anonymized all data themselves before donating it to the PRINCIPLE project in TMX format. The only additional processing that was done on their data was during the development of bespoke NMT engines. Three LRs were developed based on data Ciklopea d.o.o. donated, which are not included in further LR descriptions.

We had contacted additional public sector institutions to the ones mentioned above, but were not successful in establishing collaboration. Based on the experiences of successful and unsuccessful collaborations with the public administrations, we have identified the following main challenges in collecting data for the development of parallel corpora from such institutions. a) Identifying what public sector institutions and departments produce parallel documents in two or more languages since the majority of documentation is produced in either Croatian or in other languages directly without a Croatian equivalent. b) Pairing of parallel documents since the majority are scattered over different departments and/or on various computers. c) Bureaucracy since sharing of some documents needed to be subjected to complex internal protocols. d) Intellectual property concerns since it was unclear who was the owner of the content, specifically for translations that were outsourced. e) Privacy concerns since some data contain sensitive information, and unwillingness to share the data for anonymization. f) Shortage of manpower since consolidating the data prior to donating is time consuming and not part of the provider's regular workflow.

5. Data Cleaning and Processing

One of our aims was to normalize the collected data and to generate resources in a unified format which would be suited for MT system development. Hence, the end goal of the data cleaning and processing was a sentence-aligned corpus in TXT format[20]. Given the variety of

[18] https://marcell-project.eu

[19] The following previously developed resources have been uploaded to the ELRC-SHARE repository:
- SETimes parallel corpus (Agić and Ljubešić, 2014) https://opus.nlpl.eu/SETIMES.php
- hrenWaC (Ljubešić et al., 2016) https://opus.nlpl.eu/hrenWaC.php
- hrenWaC 2.0 https://www.clarin.si/repository/xmlui/handle/11356/1058.

[20] Note that some data providers handed over data in TMX format which was already manually aligned at the TU level. Upon inspection, these units were often sentences, but sometimes also comprised smaller or greater units. We uploaded

formats the data was originally provided in, we had to implement a number of processing steps to obtain the desired format.

5.1 Text Extraction

In principle, documents formatted in HTML and MS Word format proved easier to work with in comparison with the PDF format. For the former, we inspected the contents and manually copied the text into plain text files, while making interventions on two fronts: a) when parts of a document were missing or untranslated, we removed those sections to minimize alignment problems, and b) we did not incorporate all the text from tables, picture descriptions and formulas found in documents containing annual reports or financial statements which included a large amount of numerical data with little to no useful language data.

When it comes to the PDF documents, many contained selectable text, while others were simple scans of original documents where text could not be extracted. Due to this, the extraction from PDF documents was both done manually and using optical character recognition (OCR) software[21] where needed. When manually extracting content, we followed the same guidelines as for HTML and MS Word documents. Extracting footnotes presented additional issues for both manual and OCR data extraction: as sentences do not always end at the bottom of a page, footnote extraction had to be supervised in order not to split sentences and paragraphs.

The majority of the bilingual data was provided as two separate documents, one for each language, however some documents came as two-column PDFs in which each column represents one of the languages. Naturally, we had to split these documents into parallel monolingual TXT files. The processing of these PDF documents proved to be the most time-consuming task, as no straightforward automated solution was available. In addition, some of the PDF documents were not formatted correctly, so we had to handle them with particular diligence.

5.2 Data Cleaning

Upon extracting the text into TXT files, we had to additionally clean the data to improve its quality, as some of the provided data was noisy to begin with, and using OCR introduced additional noise. Certain issues were encountered across the board: there were unnecessary parts of the text (e.g. point strings and page number tags in the content), boilerplate content was frequently repeated, sentences were broken into multiple lines, bullets into separate lines, etc. Other issues were specific to OCR: incorrect recognition of diacritics found in English texts (e.g. *Zavižan* was recognized as *Zavizan*), incorrect recognition of letters in Croatian texts (e.g. the letter *đ* was recognized as *d*), appearance of the ¬ sign inside of words, tabs instead of spaces, whitespace gaps, etc. A bigger problem was when two words were joined together, as this could not be detected automatically. This means that every document was skim-read to detect words that were joined. Additionally, words that had a footnote label next to them were often combined with the footnote number (e.g. *Hrvatske3*), and it was even more complicated when the footnote number was combined with, for example, a year or other numerical data. Tables also required special attention: each table was manually checked in case it happened to split a sentence into two parts, or in case multiple columns of one row were merged.

One of the most interesting errors that occurred only in some of the searchable PDFs are ligatures: a product of a particular text font connecting or merging two letters into one letter, resulting in spacing errors when the text is copied from the PDF. Connecting letters with ligatures usually happens when the capital 'T' is followed by the letter 'h' which appears in frequent English words such as: *Th is*, *Th ey*, *Th e*. The other common case in which ligatures appear are words with the letter 'f' followed by letters such as 'i', 'l', 'f'. Examples include: *profi l*, *fi ltar*, *jeft iniji*, *Direzione Aff ari Internazionali*. Their frequency in both Croatian and English was considerable, but the list of affected strings was finite. This allowed us to identify all ligature sequences that occur in the texts, and then group them into categories based on whether they can be corrected automatically or not, which facilitated a straightforward cleaning process. The sequences were categorized as follows: a) the sequence is a ligature in both Croatian and English documents, b) the sequence is a ligature only in Croatian documents, c) the sequence is a ligature only in English documents, and d) the sequence has to be checked manually for every single match.

Alongside PDFs, there was considerable noise in some of the HTML documents, but the errors were more often related to individual words than to structural issues. The use of regular expressions proved effective in finding the errors, but not in their automatic correction. For example, a common mistake was the combination of a dot and a word, i.e. the dot was either in front of the first letter (e.g. *.treaty*) or inside the word (e.g. *implementa.tion*). Furthermore, letters with diacritics appearing within English words, e.g. *Condžttiđžns*, were also somewhat challenging. Croatian texts exhibited similar phenomena: *meQutim* instead of *međutim*, *me8unarodne* instead of *međunarodne*, etc. The biggest problem with such errors is that not all combinations could be found and almost every document had its own specific examples.

Finally, once all cleaning was completed, we used a sentence aligner to align the parallel documents at the sentence level. Specifically, we used vecalign (Thompson and Koehn, 2019), a state of the art automatic alignment tool that uses fastText embeddings (Grave et al., 2018) to calculate the alignments of the TUs[22] in our processed corpora. In terms of parameters, we set the maximum number of allowed overlaps to 5, maximum alignment size to 4, and during embedding training we used the provided English tokenizer for the English side of the corpus, while we used the provided Slovene tokenizer for the Croatian side of the corpus, as a Croatian tokenizer was not provided in vecalign's pipeline. As expected, this

them as is, foregoing automatic alignment, as they already satisfy the required format and can be considered gold-standard.

[21] We used the commercial ABBYY Finereader OCR software. https://www.abbyy.com

[22] Note that while none of the TUs in these datasets are larger than a single sentence, they can be smaller, as they sometimes contain text segments like list entries, table cell content, section titles or subtitles, which are often not complete sentences and can be as short as a single word or phrase.

did not seem to cause any issues, likely due to the high similarity between Slovene and Croatian. After performing the alignment we manually checked a random subsample of aligned sentence pairs to confirm the tool's accuracy. On 100 randomly sampled sentence pairs, 98 were accurately aligned. This high accuracy is likely due to the fact that the parallel data was extensively preprocessed and was well-prepared for automatic alignment. Any incorrect translation pairs are more likely to be a consequence of noise in the parallel documents, rather than a mistake of the alignment tool itself.

6. Corpora Statistics

After completing the processing steps the resources were ready for publication. Here we present an overview of the 13 resources categorized by domain. Cumulative descriptive statistics per domain are provided in Table 2. All resources contain at least the Croatian-English pair.

Domain	**TUs**
eJustice	738,923 (88.71 %)
eProcurement	22,703 (2.73 %)
eHealth	563 (0.07 %)
General	70,810 (8,5 %)
Total	*832,999*

Table 2: Translation unit (TU) counts for the 13 corpora as grouped by DSI domains.

6.1 eJustice Domain

Eight resources belong to the eJustice domain, seven of which are parallel corpora, while one is a glossary of legal terms. In addition to the Croatian-English language pair present in all the resources, the glossary also contains translations in German. One of the resources has been additionally filtered and evaluated via an MT development pipeline. They were provided by 4 different data providers: Croatian Ministry of Foreign and European Affairs, the State Commission for Supervision of Public Procurement Procedures, the Central State Office for the Development of the Digital Society and the Faculty of Humanities and Social Sciences, University of Zagreb. As such, they contain a variety of legal documents, EU court judgements and international agreements and are all freely available, totalling 738,923 TUs (see Table 3).

6.2 eProcurement Domain

There are 3 parallel corpora belonging to the eProcurement domain, each donated by a different data provider: the Central State Office for the Development of the Digital Society, the State Commission for Supervision of Public Procurement Procedures and the Central Public Procurement Office. They contain a variety of public procurement documents, including directives of the European Parliament and of the Council. In total, they contain 22,703 TUs (see Table 4).

6.3 eHealth Domain

There is one parallel corpus belonging to the eHealth domain. It was donated by the Croatian Ministry of Foreign and European Affairs and contains decisions related to the COVID-19 disease pandemic. It contains 563 TUs (see Table 5).

6.4 General Domain

The remaining parallel corpus belongs to the General domain. It was donated by the Central State Office for the Development of the Digital Society and contains a wide variety of documents on a mixture of topics such as newsletters, tax regulations, science and statistical information. It contains 70,810 TUs (see Table 6).

Corpus name	**TUs**
PRINCIPLE MVEP Croatian-English-German Glossary of Legal Terms	1,485
PRINCIPLE DKOM Croatian-English Parallel Corpus of legal documents	492
PRINCIPLE MVEP Croatian-English Parallel Corpus of legal documents	113,685
PRINCIPLE MVEP Croatian-English Parallel Corpus in the legal domain (evaluated)	110,649
PRINCIPLE MVEP Croatian-English Parallel Corpus of Court Judgements	13,335
PRINCIPLE SDURDD Croatian-English Parallel Corpus in the legal domain	261,046
PRINCIPLE SDURDD Croatian-English Parallel Corpus of international agreements	234,500
PRINCIPLE FFZG Croatian-English Parallel Corpus in the eJustice domain	3,731

Table 3: Translation unit (TU) counts for the 8 resources belonging to the eJustice domain.

Corpus name	**TUs**
PRINCIPLE SDURDD Croatian-English Procurement Parallel Corpus	3,911
PRINCIPLE DKOM Croatian-English Parallel Corpus of Directives of the European Parliament and of the Council	11,511
PRINCIPLE Central Public Procurement Office of Republic of Croatia Croatian-English Procurement Parallel Corpus	7,281

Table 4: Translation unit (TU) counts for the 3 resources belonging to the eProcurement domain.

Corpus name	**TUs**
PRINCIPLE MVEP Croatian-English Parallel Corpus of Decisions related to the COVID-19 disease epidemic	563

Table 5: Translation unit (TU) counts for the resources belonging to the eHealth domain.

Corpus name	**TUs**
PRINCIPLE SDURDD Croatian-English Parallel Corpus in the General Domain	70,810

Table 6: Translation unit (TU) counts for the resources belonging to the General domain.

7. Conclusion

As a result of the CEF-funded project PRINCIPLE, a total of 20 distinct Croatian LRs have been developed: 19 parallel corpora and 1 glossary. All LRs are uploaded to the ELRC-SHARE repository under various licenses, and many are freely available. We believe we have made a substantial contribution to the improvement of the Croatian-English language pair in the eTranslations system in two DSI domains. On the ELRC-SHARE repository at the time Croatian LRs were contributed (May 2021), 5 (26 %) out of 19 Croatian LRs were in the eProcurement domain and 10 (34 %) out of 29 were in the eJustice domain. We have made a moderate contribution to the eHealth domain by uploading 3 (6 %) out of 49 Croatian LRs.

In this paper we presented 13 freely available LRs developed from data donated by six data contributors from the public administration, and presented the particular challenges associated with data collection, cleaning and processing. The LRs cover three DSI domains (eJustice, eProcurement and eHealth) as well as data in the general domain, sizing in total 832,999 TUs. In order to continuously collect public administration data and develop LRs from this data, it would be beneficial for the Croatian language to have data donation processes incorporated into workflows of data creators. However, language data collection has not been identified as a priority in Croatia, as there is no infrastructure or (financial) support on the national level that would serve as a hub for collection and processing of language resources and tools as well as a center for educating stakeholders interested in contributing, developing and/or using Croatian language technologies.

8. Acknowledgements

PRINCIPLE was generously co-financed by the European Union Connecting Europe Facility under Action 2018-EU-IA-0050 with grant agreement INEA/CEF/ICT/A2018/1761837. We wish to thank all our data providers for donating their valuable time and data, as well as all our colleagues and students who contributed to the project. This paper was written with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at Technological University Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme. We thank the anonymous reviewers for their insightful comments.

9. Bibliographical References

Agić, Ž and Ljubešić, N. (2014). The SETimes.HR Linguistically Annotated Corpus of Croatian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC' 2014)* (pp. 1724-1727).

Ghaddar, A., & Langlais, P. (2020). SEDAR: a Large Scale French-English Financial Domain Parallel Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 3595-3602).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 3483-3487).

Ljubešić, N., Esplà-Gomis, M, Toral, A.,Ortiz Rojas, S., Klubička, F. Producing Monolingual and Parallel Web Corpora at the Same Time - SpiderLing and Bitextor's Love Affair. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2949-2956).

Moran, R., Escartín, C. P., Ramesh, A., Sheridan, P., Dunne, J., Gaspari, F., Castilho, S., Resende, N. and Way, A. (2021). Building MT systems in low resourced languages for Public Sector users in Croatia, Iceland, Ireland, and Norway. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track* (pp. 353-381).

Rehm, G., Marheinecke, K., Hegele, S., Piperidis, S., Bontcheva, K., Hajič, J., Choukri, K., Vasiļjevs, A., Backfried, G., Prinz, C., Gómez Pérez, J. M., Meertens, L., Lukowicz, P., van Genabith, J., Lösch, A., Slusallek, P., Irgens, M., Gatellier, P., Köhler, J., Le Bars, L., Anastasiou, D., Auksoriūtė, A., Bel, N., Branco, A., Budin, G., Daelemans, W., De Smedt, K., Garabík, R., Gavriilidou, M., Gromann, D., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Odijk, J., Ogrodniczuk, M., Rögnvaldsson, E., Rosner, M., Sandford Pedersen, B., Skadiņa, I., Tadić, M., Tufiş, D., Váradi, T., Vider, K., Way, A., and Yvon, F. (2020). The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)* (pp. 3322-3332).

Rehm, G. and Uszkoreit, H., editors. (2012). META-NET White Paper Series: Key Results and Cross-Language Comparison. *META-NET White Paper Series. Kaiserslautern: Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI). https://web. archive. org/web/20181219124131/http://www. meta-net. eu/whitepapers/key-results-and-cross-language-comparison (13.12. 2018)*.

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., and Váradi, T. (2014). An update and extension of the META-NET Study "Europe's Languages in the digital age".

Thompson, B. and Koehn, P., 2019, November. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1342-1348).

Utka, Andrius; Rackevičienė, Sigita; Rokas, Aivaras; Bielinskienė, Agnė; Mockienė, Liudmila and Laurinaitis, Marius, 2022, *English-Lithuanian Comparable Cybersecurity Corpus - DVITAS*, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/47.

Váradi, T., Koeva, S., Yalamov, M., Tadić, M., Sass, B., & Nitoń, B. (2020). The MARCELL Legislative Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2020)* (pp. 3761-3768).

Way, A, Bago, P, Dunne, J., Gaspari, F., Kåsen, A., Kristmannson, G., McHugh, H., Olsen, J. A., Sheridan, D. D., Sheridan, P., Tinsley, J. (2020.) Progress of the PRINCIPLE Project: Promoting MT for Croatian, Icelandic, Irish and Norwegian. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 465-466).

Way, A., Gaspari, F. (2019). PRINCIPLE: Providing Resources in Irish, Norwegian, Croatian and Icelandic for the Purposes of Language Engineering. In *Proceedings of MT Summit XVII, volume 2* (pp. 112-113).