information

Article

Four Million Segments and Counting: Building an English-Croatian Parallel Corpus through Crowdsourcing Using a Novel Gamification-Based Platform

Rafał Jaworski [1,*], Sanja Seljan [2] and Ivan Dunđer [2]

[1] Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznań, 61-712 Poznań, Poland
[2] Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, 10000 Zagreb, Croatia; sanja.seljan@ffzg.hr (S.S.); idundjer@ffzg.hr (I.D.)
* Correspondence: rjawor@amu.edu.pl

Abstract: Parallel corpora have been widely used in the fields of natural language processing and translation as they provide crucial multilingual information. They are used to train machine translation systems, compile dictionaries, or generate inter-language word embeddings. There are many corpora available publicly; however, support for some languages is still limited. In this paper, the authors present a framework for collecting, organizing, and storing corpora. The solution was originally designed to obtain data for less-resourced languages, but it proved to work very well for the collection of high-value domain-specific corpora. The scenario is based on the collective work of a group of people who are motivated by the means of gamification. The rules of the game motivate the participants to submit large resources, and a peer-review process ensures quality. More than four million translated segments have been collected so far.

Keywords: parallel corpus; data acquisition; gamification; crowdsourcing; machine translation; natural language processing

Citation: Jaworski, R.; Seljan, S.; Dunđer, I. Four Million Segments and Counting: Building an English-Croatian Parallel Corpus through Crowdsourcing Using a Novel Gamification-Based Platform. *Information* **2023**, *14*, 226. https://doi.org/10.3390/info14040226

Academic Editor: Marjan Mernik

Received: 14 February 2023
Revised: 23 March 2023
Accepted: 2 April 2023
Published: 6 April 2023

1. Introduction

The use of machine translation services has nowadays become a standard way for acquiring and comprehending information and data that are written in foreign languages. In a globalized world with more than 7000 languages [1], multilingual communication is essential regardless of the type of business, research, education, etc. Therefore, building language resources and tools, such as digital corpora and machine translation systems, which can be used independently or be integrated into other tools such as Computer-Assisted Translation (CAT) tools, represent an important element in business and research.

It is estimated that 50% of all languages are low-resourced, although this term encompasses various definitions, including being a language with limited language resources and rarely used in language technologies, having a limited number of labeled datasets, having a limited online presence, having a small number of speakers, etc. [2]. Even if parallel data exists, it is oftentimes of lower quality or originates from very specific sources, such as religious texts or IT documentation, which are usually very different from the desired domain. The domain, however, is crucial for the implementation of effective machine translation systems. Therefore, the lack of data, the low data quality, and the noisiness of the data are common problems for many languages.

Parallel corpora represent a fundamental resource for many different research tasks and scientific analyses, for building various applications, and for educational purposes. Parallel corpora are used as an indispensable source in the field of Natural Language Processing (NLP) [3]. This is related to building machine translation systems and creating translation memories that are commonly used in Computer-Assisted Translation (CAT) tools and Named Entity Recognition (NER) systems for building dictionaries, text mining, and extracting collocations [4]. Software that is based on parallel corpora, such as concordance searching applications [5], machine translation systems, and CAT tools, directly depend on the quality of parallel corpora, their size, domain, and language pair.

Research conducted by [6] shows that translators who work in a real translation environment and use machine translation systems generally have benefits in terms of productivity with regards to the use of machine translation systems.

These systems have been analyzed by applying several quality assurance and evaluation methods, as in [7], where the authors performed an extensive quality assessment of parallel resources used in CAT tools, or by using automatic quality metrics for evaluating machine translation systems [8–11].

Parallel corpora can be used in the education process, especially in the domain of computer and information sciences, for research on NLP [12], in language studies on tasks of evaluating and assessing semantics [13], for the translation process and terminology analysis [14], for post-editing tasks after the use of machine translation [15], or CAT technology [16]. However, building scalable, high-quality parallel corpora is a challenging and resource-intensive task in terms of time, effort, cost, and knowledge.

One of the main issues with statistical machine translation (SMT) and neural machine translation (NMT), which have become dominant approaches for building machine translation systems, is the lack of large-scale parallel data, which is especially relevant for low-resource languages [17–19].

SMT relies on statistical models that use parallel corpora to identify patterns and relationships between words in the source and target languages. This information is then used to translate new sentences [20].

NMT, on the other hand, uses deep learning techniques to generate translations [21]. An NMT system is trained on large amounts of bilingual data and intends to learn a shared representation of the source and target languages that can be used for translation.

As one of the 24 official EU languages, Croatian still suffers from a significant lack of bilingual data and has limited data resources available, which are necessary for developing a variety of language technologies, including machine translation systems.

In order to facilitate the acquisition and management of parallel corpora, it is desirable to have a digital platform that is language independent, easy to use, and accessible to a large number of users. For this purpose, a web-based application called “TMrepository” was designed in an effort to provide a straightforward, customizable, and free service needed for collecting and storing parallel corpora. The platform is based on the concepts of crowdsourcing and gamification, which make the tedious task of collecting parallel data more effective, appealing, and pleasant.

The main goal of this paper is to present an English-Croatian parallel corpus that was created by using a specially built web-based platform. Furthermore, the specific aims of this paper are as follows:

(i) To analyze the importance of parallel corpora, specifically for machine translation purposes.
(ii) To present the integration of crowdsourcing and gamification methods into a new web-based platform for creating and organizing parallel corpora.
(iii) To demonstrate the functionalities of the created system.
(iv) To analyze the resulting English-Croatian parallel corpus that contains more than four million segments, i.e., more precisely, translation units.

It should also be noted that segments can be understood as text chunks, i.e., text lines that do not necessarily end with a sentence delimiter. In a corpus, they are fundamental logical units that have the tendency to be repetitive, and they come in the form of a whole sentence, parts of sentences, multi-word units, phrases, or even abbreviations. A typical parallel corpus consists of corresponding pairs of segments in the source and target languages. Since they are stored line by line in a corpus, they can be treated as translation units.

This paper is organized in the following way: In the Introduction section, the motivation for building a web-based platform for collecting and storing parallel corpora is discussed. The second section presents related work and research on building parallel corpora for machine translation, crowdsourcing, and gamification. In the third section, the crowdsourcing platform of "TMrepository" is presented, along with its main functionalities. The fourth section exhibits details of the experimental setup, and the research results are elaborated in terms of the harvested parallel corpus, which contains more than four million parallel segments. Finally, in the last section, conclusions are stated and suggestions for further research are given.

2. Related Work

2.1. Building Parallel Corpora for Machine Translation Systems and for Low-Resourced Languages

Building parallel corpora is essential for all sorts of language-oriented analyses, for building various tools and different types of resources. This is especially true for low-resourced languages that are characterized by scarce monolingual and bilingual data, which is increasingly important in today's multilingual global communication.

The two dominant types of machine translation system architectures are statistical machine translation (SMT) and neural machine translation (NMT), but they differ in their approach to translation. NMT in particular is currently a hot topic of interest to researchers, engineers, industry stakeholders, and language specialists. According to [22], NMT has become the dominant approach for building machine translation systems in which artificial neural networks are utilized.

Statistical Machine Translation (SMT) is typically understood as an approach that applies phrases that are extracted during model training and are not (necessarily) linguistically motivated [20]. This approach relies on model features that are trained separately and later combined in a machine translation system implementation [23]. A basic SMT model consists of a language model that covers the target language and is hence trained on monolingual data, a translation model that is trained on parallel data and that stores key statistical information on word occurrences and corresponding translations, and a decoder that handles the actual translation.

Similar to SMT, NMT uses large amounts of parallel data [17] for model training; however, it utilizes data-driven deep learning methods in the machine translation process [21]. It applies artificial neural networks for predicting word sequences and does not train model features separately. NMT models are typically based on the encoder-decoder architecture. The encoder handles the source text and converts it into a continuous hidden representation, whereas the decoder generates the target text conditioned on the hidden representation generated by the encoder. The encoder and decoder are typically implemented as recurrent neural networks (RNN) [24], convolutional neural networks (CNN) [25], or transformer networks [26].

Overall, NMT is regarded as a more advanced method of machine translation and is capable of producing translations that are more complex, fluent, and accurate, as well as ones that better capture semantic meaning and contextual information. However, SMT is significantly less computationally expensive and thus still has a place in certain applications where speed and efficiency are a top priority.

Nevertheless, both architectures rely on high-quality in-domain parallel corpora, which poses a challenge for low-resource languages. One research [27] presented the process of building a parallel corpus consisting of more than 10 thousand segments in order to build an SMT system for English-Manipuri (Indian language).

Another extensive study [28] analyzed domain adaptation techniques and the impact of general-domain and industry-related parallel corpora on building phrase-based SMT systems for the Croatian language.

The process of corpus collection, expanded by scraping websites or by applying Optical Character Recognition (OCR), was presented in [29]. The corpus of a low-resource Indian language (Odi) and English resulted in 98,302 parallel segments, which derived from various domains, such as religious texts, literature, government policies, everyday communication, the general domain (Wikipedia), etc. The corpus was collected for the purpose of building a machine translation system that would be used by the research community.

A paper by [30] examined the use of Amazon Mechanical Turk, a crowdsourcing platform that was utilized for creating parallel corpora, and addressed the different challenges when collecting resources.

A process of compiling multilingual parallel corpora for Uzbek, Russian, and English by using a CAT tool as a platform for collaborative work is shown in [31].

One study presented methods for collecting Turkish-English parallel corpora by crawling a journal website with 6500 Turkish abstracts and their corresponding translations into English, then converting them into a translation memory that was used in a CAT tool, and then further for extracting terminology in the medical domain and for building a machine translation system [32].

Data augmentation strategies in SMT, such as appending entries from multilingual dictionaries directly to the bitext, and substituting glosses in place of complex inflected forms in the source language, have been demonstrated in [33]. This was performed in order to create new source data that is more similar to the target language.

The use of parallel corpora and the rule-based approach for building translation systems for low-resourced languages and dialects of Spain (Aranese Occitan, Aragonese, and Asturian/Leonese) have been emphasized in [34]. The rule-based approach to machine translation relies solely on linguistic information about the source and target languages, which is usually retrieved from dictionaries, grammars, and other linguistic resources that cover the semantics, morphology, and syntactic regularities of each involved language.

When it comes to parallel corpora for NMT, it is discussed in numerous studies with regard to corpus size, domain adaptation [35], multilingual translation [36], specific domain problems [37], automatic and human evaluation methods, etc.

According to [17], NMT suffers heavily from the high cost of collecting large-scale parallel data. For this reason, many studies on low-resource languages were performed using very limited amounts of data, either by using data from auxiliary languages with similar syntax and semantics or by using multimodal data that combined text and images.

In one study, the authors created corpora for low-resource languages (Gujarati, Kazakh, and Somali) that are used for building NMT systems by adding comparable data and a bilingual dictionary [38].

In addition to augmenting the original training data with parallel phrases extracted from the original training data using a statistical machine translation system, an NMT system using Gated Recurrent Unit (GRU) and transformer networks for Hindi-English and Hindi-Bengali has also been proposed [39].

Other authors [40] used naïve regularization methods in NMT, based on sentence length, word frequencies, and punctuation, in order to penalize translations that are very different from the input sentences, and this approach proved to consistently enhance translation quality across multiple low-resource languages with varying training data sizes.

Challenges in building NMT systems in terms of parallel corpora, such as domain mismatches, amount of training data, rare words, size of sentences, word alignments, etc., have been analyzed in [41]. In comparison to the SMT model, the authors reported that NMT had lower quality for out-of-domain translations, especially for the criterion of adequacy, as also confirmed in a study by [42]. NMT systems perform better when large amounts of parallel data are available, i.e., worse for low-resource language pairs. However, NMT systems perform better for extremely rare words. The same authors [41] also report problems with longer sentences in parallel corpora (i.e., longer than 60 words), as the attention mechanism in the model does not always perform well with regard to word alignment and beam search decoding. The corpora in this research consisted of several domain-specific subcorpora ranging from less than 340,000 segments up to almost 14,000,000 segments. The authors reported similar results for SMT and NMT for in-domain training—NMT performed better for the domains of IT and subtitles, whereas SMT was better for domains such as law, medicine, and religious texts. Output for out-of-domain data was worse for NMT.

Challenges in collecting English-Ethiopian parallel corpora in the domain of religion, where the NMT efficacy increased with the availability and number of parallel datasets, have been presented in [43]. Datasets were collected for four Ethiopian languages, one having almost 27,000 segments and the others having less than 8000 parallel segments.

Another study [18] performed research on NMT for two low-resource language pairs—French-Vietnamese and English-Vietnamese—and used two methods to improve the translation of rare words. The first one uses dynamical learning of word similarity of tokens in the shared space among source languages, whereas another one attempts to augment the translation ability of rare words through updating their embeddings during the training. Here, the parallel data was obtained from TED Talks and consisted of 231,000 English-Vietnamese and 203,000 French-Vietnamese sentences. To generate synthetic bilingual data, the authors sampled 1.2 million English monolingual sentences from the European Parliament's English-French corpus.

A detailed study on existing research advancements in the low-resource language NMT was conducted by [19]. The paper included major NMT techniques applicable to low-resource language pairs and provided a holistic overview of the entire research landscape with future directions on how to increase research efforts. Due to the evident lack of parallel data, the researchers used various techniques for creating additional resources, such as adding data from different resources by using bilingual dictionaries, back-translation, monolingual data selection, or parallel corpus mining from comparable corpora with sentence ranking. This analysis revealed that data augmentation methods have recently attracted a lot of interest in the research community.

A publicly available Japanese-Chinese corpus that focuses on spoken language data consists of approximately 1.4 million sentence pairs of bilingual data. It was constructed through a large-scale collection of Japanese-Chinese bilingual sentences from subtitles, which were manually aligned, evaluated, and tested in different translation experiments [44].

Multilanguage NMT models can be used when machine translation needs to be performed among more than one language pair [36,45]. This is especially interesting when a smaller number of languages is present in the machine translation system and when the system needs to generate translations between closely related languages [46,47].

NMT can also employ transfer learning. Here, a parent model is first trained on a large corpus of parallel data from a high-resource language pair, which is then used to initialize the parameters of a child model that is trained on a smaller parallel corpus [48–51].

Zero-shot NMT is used when there is no available parallel data for a specific language pair. The model in this approach is trained with no parallel data for the considered language pair, i.e., by using an intermediate pivot high-resource language pair [52]. Here, the translation process is decomposed into two training phases: source-pivot and pivot-target language pairs [53]. This approach, however, is very time-consuming and suffers from backpropagation errors, as inaccuracies from the first training phase are transferred to the second training phase. In order to reduce these problems, it is possible to allow interaction between these two models by sharing word embeddings of the pivot language or combining pivoting with transfer learning. Each approach can make use of submodels and techniques, which can be combined to create the resources needed for low-resource language pairs, improve output quality, and reduce time, effort, and costs.

2.2. *Crowdsourcing and Corpora Acquisition*

Crowdsourcing is the process of recruiting undefined, very often previously unknown, and numerous individuals by stating an open call in order to complete a certain task that would otherwise be given to stakeholders, in-house personnel, etc. [54].

The labor of a large, talented, and interested group is not free or cheap, but crowdsourcing costs significantly less than what regular internal employees are paid [55]. It enables the crowd to carry out distinct tasks that were once only solvable by a small, highly specialized group of individuals [56].

A study by [57] emphasizes several important aspects of crowdsourcing, such as motivation, quality, aggregation, human skills, participation time, and cognitive load. The paper classifies crowdsourcing into seven genres: games with a specific purpose in mind, mechanized labor with subsequent payment, the wisdom of crowds with a large number of people participating and thinking independently, crowdsourcing by the unpaid general public that is motivated by curiosity, dual-purpose work as a means to perform a task that cannot be performed automatically, grand search with the task of finding a solution for a specific problem, and knowledge collection from volunteer contributors based on the idea to create large databases of common facts.

One corpus-related paper [58] presents crowdsourcing that is used for the acquisition of annotated corpora, which are essential for various NLP tasks and executing algorithms. Here, crowdsourcing is approached using a project-based strategy with distinct phases.

Crowdsourcing can be used for collecting training data and in order to perform data annotation, clustering, and parsing, supported by a statistically based NLP analysis [59]. Nevertheless, according to the same authors, the use of crowdsourcing in scientific studies is still relatively new.

Crowdsourcing can also be utilized for the collection of unstructured documents and reports and for open big data analysis [60]. In this research, multiple methods have been implemented, such as search-log-based detection, machine learning, and crowdsourced annotation, with the aim of detecting seasonal medical events that happen globally.

Today, crowdsourcing is also used to help meet the demand for translation services, due to the fact that collaborating on translation projects clearly has numerous benefits [54]. Another study discussed the vast potential of collaborative work for different NLP tasks [61].

A multiphase workflow for language translation has been shown to be a very cost-effective crowdsourcing approach [62]. Here the task of sentence translation is carried out in three phases: word translation, assisted sentence translation, and synthesis of translations. The advantage of using such subtasks is that everything is consistent and verifiable at all times, which leads to better translations for a variety of diverse and non-expert translators.

A paper confirmed that crowdsourcing can contribute significantly to the creation of essential language resources [63]. The authors used Amazon Mechanical Turk and showed that it is indeed possible to obtain translations of high quality from non-professional translators while at the same time keeping the overall costs below the price of professional translations.

Another research paper discussed the various problems and different limitations when preparing a crowdsourced translation task on Amazon Mechanical Turk [30]. Specifically, when working with non-professional translators, quality concerns must be addressed, e.g., in order to detect any illicit use of web-based and freely available automatic translation services.

Some other authors considered crowdsourcing translations to be the translation industry's next technological breakthrough [64]. Interactive and scalable machine translation that helps the crowd and maximizes its potential will be crucial, especially in terms of translation or post-editing, whereas the translation environment here becomes capable of expanding its ability to handle massive amounts of data.

When it comes to education, teaching computer-assisted translation today is inextricably linked to the development of new translation standards, novel approaches, and new technologies in the translation industry [65]. According to the same research, CAT depends on collaborative translation, machine translation, translation management systems, and especially crowdsourcing.

Distance and blended learning are also important features of the training and teaching environment for NLP-related tasks [66]. Classroom facilities limit teaching capabilities and educational approaches, whereas available hardware and its diversity must also be considered. Furthermore, different operating systems, available software, and other variables also have an impact on what is considered to be the best fit for particular training objectives.

This is why a supportive language technology platform is proposed in this paper. It comes in the form of a cloud-based, open-source corpus management system that adopts the crowdsourcing concept, which has shown to be useful for teaching purposes and the acquisition of various language resources. The idea behind this online system is to enable students with internet connectivity to access translation technology from anywhere and, at the same time, to reduce teaching costs. Such a system should also enable students to participate in real-time collaborative work and many exercises, such as NLP tasks and post-editing of machine translations, etc., while also allowing them to submit translation-specific assignments and track their own exercise progress and pace of study.

2.3. The Role of Gamification and Its Potential in NLP

According to [67], gamification can be understood as the application of game design principles to change behavior in a non-gaming context. It can boost participants' involvement and the level of interaction between various elements in a given environment, given that it is implemented and applied properly.

Gamification primarily tends to increase the positive motivation of users towards specific activities or the use of technology in a game-like scenario. This also increases the output or outcome of specific activities in terms of both quantity and quality [68]. Still, applying gamification is not always straightforward, as various tasks cannot be easily reduced to the level of games. In other words, their complexity makes it difficult to incorporate gamification in other, i.e., non-gaming environments.

Gamification also involves, to some extent, the understanding and deeper knowledge of human psychology, as it usually aims to affect users' behavior, and this also increases the difficulty of designing and applying gamification-based environments [68].

According to one study, gamification has been successfully used to verify machine translation quality, which is crucial in the performance assessment of machine translation systems. Here the intention of the gamification strategy was to keep the evaluators engaged by continually asking them to provide a quality score for a machine translation, and in return they received feedback and rewards in the form of stars that reflected how close their score was to a reference, i.e., the gold-standard score [69].

An efficient approach to gamification in natural language processing is found in [70]. The paper describes the platform Gonito.net, which derives its name from the Polish name of the game "tag". Its main purpose is to foster competition in NLP-related tasks among a group of researchers. Each of them is given a specific problem to solve, such as predicting the year of publication of a Polish text, predicting the gender of the author of the text, or searching for legal clauses by analogy. These problems are referred to as "challenges" on the platform. The participants of a challenge receive the training, development, and testing sets, and their goal is to provide the results for the given test set. Importantly, each participant can submit the results multiple times. The current ranking is always visible to all participants, which provides good motivation. Participants are encouraged to share the details of their implementation to inspire others. It is not considered plagiarism to use the code of other researchers and improve it. The Gonito.net platform has hosted numerous challenges and made a significant step in NLP research—all thanks to the gamification approach. Another example of the use of gamification in natural language processing is the Kaggle.com platform. This platform is known worldwide, and it was one of the first to introduce the idea of open challenges for multiple participants (Gonito.net adopted this idea later) [70]. On Kaggle, it is possible to take part in a wide variety of tasks related to statistics, machine learning, and natural language processing. The datasets are distributed to all participants, who compete to produce the best possible results. Some of the challenges involve significant monetary prizes to further motivate the participants. Those prizes are typically funded by companies that see particular practical applications of the challenge. It is also not uncommon for the companies to employ the winners of the challenge. Through its gamification-oriented idea, Kaggle has furthered an enormous effort in natural language processing by fostering research and providing vast amounts of valuable datasets.

A summarized table of relevant related work that shows the different concepts, outcomes, and ideas for this paper and future research is given below (Table 1).

Table 1. Relevant work presenting key concepts, outcomes, and ideas for this and future research.

Key concept:	Parallel data acquisition and corpus compilation	
	Author(s)	Outcomes and ideas for this and future research
	Parida et al. (2018) [29]	Scraping of websites or applying Optical Character Recognition (OCR) for low-resource languages.
	Abdurakhmonova (2020) [31]	Compiling multilingual parallel corpora for Uzbek, Russian, and English using a CAT tool.
	Shearing et al. (2018) [33]	Applying data augmentation strategies in order to create parallel resources.
	Kuwanto et al. (2021) [38]	Creating corpora for low-resource languages by adding comparable data and a bilingual dictionary.
	Sen et al. (2020) [39]	Applying augmentation of data with parallel phrases extracted from the original training data for low-resource language pairs.
	Beloucif et al. (2019) [40]	Penalizing translations that are very different from the input sentences, which has shown to consistently enhance translation quality across multiple low-resource languages with varying training data sizes.
	Lambebo et al. (2021) [43]	Analyzing challenges in collecting parallel corpora for a low-resource language pair in the domain of religion and discussing the importance of the availability and number of parallel datasets.
	Ranathunga et al. (2023) [19]	Augmenting data by using bilingual dictionaries, back-translation, monolingual data selection, or parallel corpus mining from comparable corpora with sentence ranking.
	Ngo et al. (2020) [18]	Discussing how parallel data for a low-resource language pair was obtained from TED Talks and how it was prepared for machine translation.
	Singh (2012) [27]	Gathering data for a low-resource language pair in order to acquire parallel segments that were needed for building an SMT system.
	Zhang et al. (2023) [44]	Constructing a large-scale collection of bilingual sentences for a low-resource language pair from subtitles that were manually aligned, evaluated, and tested in different translation experiments.
Key concept:	Resources for building machine translation systems and their characteristics	
	Author(s)	Outcomes and ideas for this and future research
	Bahdanau et al. (2015) [22]	Discussing how NMT is used as the dominant approach for building machine translation systems that rely on large amounts of corpora.
	Koehn et al. (2003) [20]	Presenting a machine translation model that consists of different submodels that are trained separately with bilingual and monolingual data.
	Koehn (2010) [23]	Discussing how the translation model in SMT is trained on large amounts of parallel data and then tuned with additional data, whereas the language model is built with monolingual data.
	Wang, W. et al. (2021) [52]	Discussing the zero-shot approach in NMT when no parallel data is available.
	Currey and Heafield (2019) [53]	Emphasizing that NMT depends on large amounts of parallel data. However, when no parallel data is available, pivot languages can be applied.
	Kamath et al. (2019) [21]	Explaining the architecture and the use of artificial neural networks for predicting word sequences based on corpora in the NMT model.
	Dong et al. (2015) [24]	Investigating the problem of how to translate one source language into several different target languages within a unified translation model that is based on the encoder-decoder architecture.

Table 1. *Cont.*

	Gehring et al. (2017) [25]	Discussing a fast and simple architecture based on a succession of convolutional layers, in contrast to bi-directional LSTMs that are otherwise regularly being used in order to encode the source sentence.
	Vaswani et al. (2017) [26]	Proposing a new simple network architecture—the transformer—that is based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.
	Wang, R. et al. (2021) [17]	Providing an extensive survey for low-resource NMT and analyzing related works with regard to various auxiliary data sources.
	Doğru et al. (2018) [32]	Discussing parallel corpora preparation for machine translation for a low-resource language pair in the domain of medicine.
	Dunđer (2015) [28]	Implementing domain adaptation techniques and analyzing the impact of general-domain and industry-related parallel corpora on the effectiveness of SMT.
	Parida et al. (2018) [29]	Scraping of websites or applying Optical Character Recognition (OCR) for low-resource languages. The corpus was collected for the purpose of building a machine translation system.
	Forcada (2021) [34]	Discussing the use of parallel corpora and the rule-based machine translation approach for low-resource languages and dialects of Spain.
	Maruf et al. (2021) [37]	Discussing document-level NMT and assessing domain-related problems with corpora.
	Chu and Wang (2018) [35]	Analyzing corpus size and domain adaptation in the machine translation system.
	Koehn and Knowles (2017) [41]	Presenting challenges in building NMT systems in terms of parallel corpora, such as domain mismatches, amount of training data, rare words, the long sizes of sentences, word alignments, etc.
	Seljan et al. (2020) [42]	Discussing human quality evaluation results and different criteria with regard to in-domain parallel data used in machine translation.
	Dabre et al. (2020) [36]	Presenting the use of NMT when machine translation needs to be performed among more than one language pair.
	Ha et al. (2016) [45]	Discussing the encoder-decoder architecture in a multilingual NMT model.
	Lakew et al. (2018) [46]	Comparing the transformer and recurrent neural networks in a multilingual NMT environment.
	Tan et al. (2019) [47]	Using NMT with knowledge distillation in order to boost the accuracy of multilingual machine translation.
	Fikri Aji et al. (2020) [48]	Studying transfer learning in NMT, as it has been shown that it improves quality for low-resource machine translation.
	Kim et al. (2019) [49]	Exploring effective cross-lingual transfer of NMT models without using shared vocabularies.
	Dabre et al. (2017) [50]	Presenting an empirical study of language relatedness for transfer learning in NMT.
	Zoph et al. (2016) [51]	Highlighting the importance of transfer learning for low-resource NMT.
Key concept:	Crowdsourcing in data acquisition and NLP tasks	
	Author(s)	Outcomes and ideas for this and future research
	O'Brien (2011) [54]	Analyzing the benefits of using crowdsourcing in order to complete certain tasks that would otherwise be assigned to stakeholders, in-house employees, etc.
	Howe (2006) [55]	Stating that crowdsourcing costs significantly less than paying traditional in-house employees.
	Howe (2008) [56]	Using crowdsourcing to carry out distinct tasks that were once only solvable by a small, highly specialized group of individuals.

Table 1. *Cont.*

	Quinn and Bederson (2011) [57]	Presenting various aspects of crowdsourcing, such as motivation, quality, aggregation, human skills, participation time, and cognitive load.
	Sabou et al. (2014) [58]	Using crowdsourcing in project-based tasks, such as the acquisition of annotated corpora and for NLP.
	Li et al. (2015) [59]	Presenting a visualization toolkit to allow crowd-sourced workers to annotate general categories of NLP problems, such as clustering and parsing.
	Munro et al., 2012 [60]	Using crowdsourcing, NLP, and big data analysis for tracking medical events on a global scale.
	Vamshi et al. (2012) [62]	Using a collaborative workflow for crowdsourcing translation tasks.
	Zaidan and Callison-Burch (2011) [63]	Discussing how crowdsourcing can play a significant role in building necessary language resources.
	Ambati and Vogel (2010) [30]	Analyzing challenges when preparing crowdsourcing translation tasks with Amazon Mechanical Turk, especially when working with non-professional translators.
	Muntés-Mulero et al. (2012) [64]	Discussing why crowdsourcing translations will be the next big breakthrough in the translation industry.
	Muegge (2013) [65]	Emphasizing why translation technology should include collaborative translation, machine translation, translation management systems, and crowdsourcing.
	Canovas and Samson (2011) [66]	Analyzing why distance and blended learning are important features of the training and teaching environment for NLP-related tasks and what is perceived as the most suitable for specific training objectives.
Key concept:	Gamification and its potential in NLP	
	Author(s)	Outcomes and ideas for this and future research
	Robson et al. (2016) [67]	Discussing the application of game design principles in order to change behavior in a non-gaming context, increase engagement of participants, and increase the positive motivation of users towards specific activities.
	Morschheuser et al. (2017) [68]	Emphasizing why gamification tends to increase the positive motivation of users and why it increases the quantity and quality of the output or outcome of given activities.
	Abdelali et al. (2016) [69]	Stating that gamification has been successfully applied for the purpose of machine translation evaluation, by motivating participants with feedback in the form of stars.
	Graliński et al. (2016) [70]	Presenting an efficient approach to applying gamification to NLP tasks using a platform that uses a system of rewards, feedback, and reproducibility; and highlighting the differences when compared to gamification-oriented elements in Kaggle.

3. Novel Crowdsourcing and Gamification-Based Corpus Management Platform

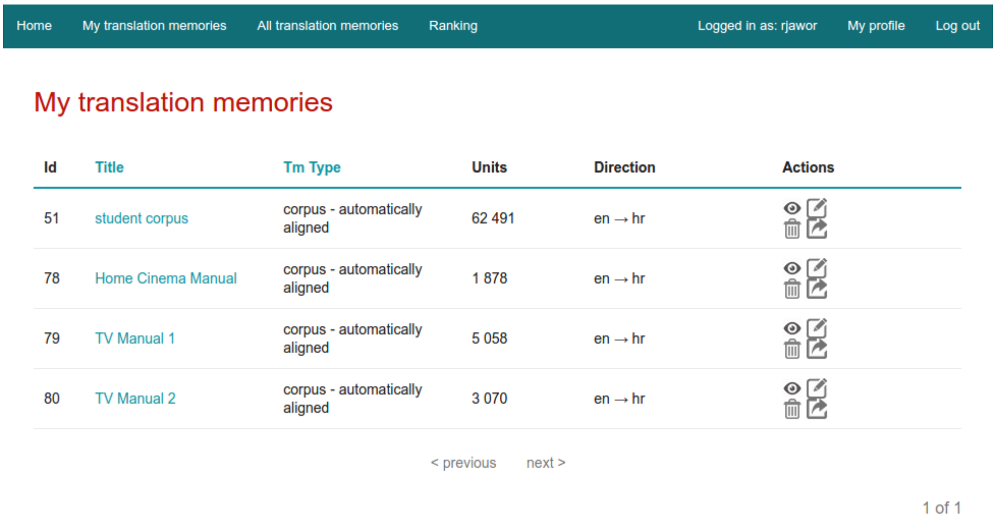
This research focuses on building an English-Croatian parallel corpus using the crowdsourcing approach and a novel gamification-based platform called "TMrepository". It is a unique online application developed with the objective of collecting parallel data, which is needed for various analyses and developing machine translation systems and new CAT and NLP tools.

Although it was mainly built with the Croatian-English language pair in mind, it can be applied to other languages as well. It is publicly available at http://concordia.vm.wmi.amu.edu.pl/tmrepository/ (accessed on 2 February 2023), and its primary purpose is to provide a user-friendly, sentence-level repository of translation memories. In this particular research, it was set up to store and manage Croatian-English and English-Croatian translation memories.

"TMrepository" additionally has gamification-inspired elements to increase the motivation of contributors—mostly students. This platform is intended to be used in the future by computer science and information science students and researchers who work with machine translations and other NLP-related tasks, or students of translation studies who create their own translation memories during their studies.

In order to attract a larger crowd, registration for the system is free and open to anyone. The user is shown a list of their own contributions after logging in (Figure 1). Uploading new resources to the system is the user's main activity, which is performed by means of an upload form. Here the user is asked to provide relevant information, such as the title of the translation memory, a brief description, and the type of resource. Available types are as follows:

- manual translation,
- manual translation—automatically aligned,
- corpus,
- corpus—automatically aligned.



My translation memories

Id	Title	Tm Type	Units	Direction	Actions
51	student corpus	corpus - automatically aligned	62 491	en → hr	👁 ✎ 🗑 ↗
78	Home Cinema Manual	corpus - automatically aligned	1 878	en → hr	👁 ✎ 🗑 ↗
79	TV Manual 1	corpus - automatically aligned	5 058	en → hr	👁 ✎ 🗑 ↗
80	TV Manual 2	corpus - automatically aligned	3 070	en → hr	👁 ✎ 🗑 ↗

< previous next >

1 of 1

Figure 1. The main window of the system—a list of uploaded translation memories.

The distinction between options "manual translation" and "corpus" depends on the author of the translation: the first option indicates that the translations were performed by the contributing user or collected from public sources (e.g., the web), whereas "corpus" indicates use of already existing translations.

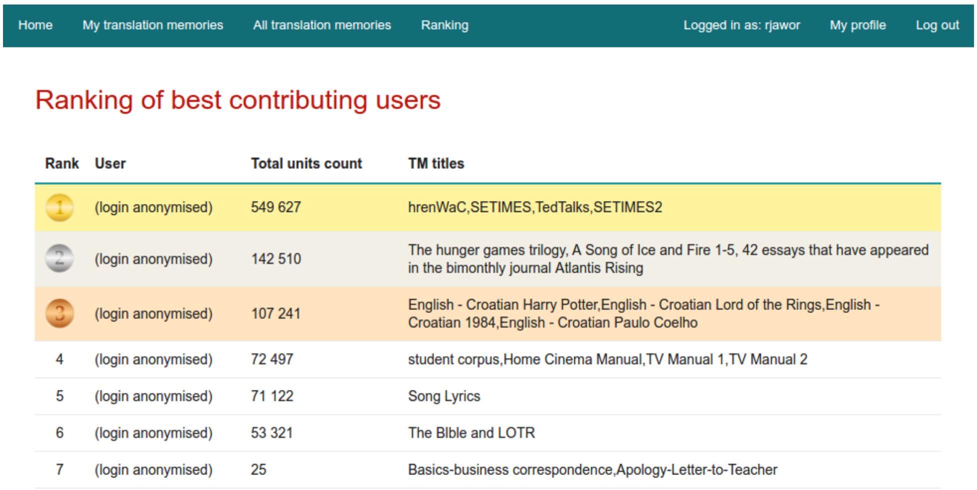
The label "automatically aligned" refers to resources that were already aligned with appropriate software (e.g., hunalign) before a user initiated an upload, as opposed to resources that are already pre-aligned with perfect or nearly perfect alignment quality.

This study focuses mainly on collecting translation memories of the type "corpus—automatically aligned", which is understood as resources automatically aligned by the contributing user and containing translations completed by other people. However, exceptions to this are possible.

Available import formats include a pair of text files, a TMX file, and a pair of Word documents. Import from TXT files assumes that two text files with UTF-8 encoding are provided, each having an equal number of lines, where one of the files contains sentences in L1 and the other in L2. Alternatively, the system is able to import TMX files. A custom-built stream TMX parser was developed to prevent problems with large TMX files using a lot of memory. The last option, a pair of Word documents in DOC or DOCX format, automatically aligns any two Word documents on the sentence level. There are no assumptions about how documents should be formatted.

"TMrepository" automatically extracts text out of uploaded documents, splits them into sentences, and performs automatic alignment with the use of the hunalign algorithm. This algorithm does not require any linguistic resources to perform the alignment. It operates in two phases: In the first pass, the source and target files are analyzed, and a rudimentary bilingual dictionary is created. Then, in the second pass, the dictionary is used to calculate the best sentence matches between the source and target sentences.

The ranking page lists all contributing users, sorted by the total number of sentence pairs uploaded in descending order. The first three users receive virtual medals and are graphically exposed (Figure 2). This gamification element is used to introduce competitiveness among the users and thus increase their motivation.



Ranking of best contributing users

Rank	User	Total units count	TM titles
1	(login anonymised)	549 627	hrenWaC,SETIMES,TedTalks,SETIMES2
2	(login anonymised)	142 510	The hunger games trilogy, A Song of Ice and Fire 1-5, 42 essays that have appeared in the bimonthly journal Atlantis Rising
3	(login anonymised)	107 241	English - Croatian Harry Potter,English - Croatian Lord of the Rings,English - Croatian 1984,English - Croatian Paulo Coelho
4	(login anonymised)	72 497	student corpus,Home Cinema Manual,TV Manual 1,TV Manual 2
5	(login anonymised)	71 122	Song Lyrics
6	(login anonymised)	53 321	The BIble and LOTR
7	(login anonymised)	25	Basics-business correspondence,Apology-Letter-to-Teacher

Figure 2. List of the best contributors.

The rankings are valuable since they also list the names of the translation memories that the top three most productive users provided. This is being performed in an effort to serve as inspiration for other users when considering suitable corpora sources. For instance, noticing that people are uploading “Harry Potter” and other books might direct the corpus search to different book titles. Similarly, the fact that one user uploaded TV manuals may encourage other contributors to search for additional translated user manuals and technical documents. Every user has access to the most recent ranking at all times.

The resources collected on the platform can be exported to various popular formats, including:

- TXT,
- TMX,
- Moses parallel files (a format widely used in machine translation system training).

The total size of the resources collected in the TMX format exceeds 200 MB. However, it is important to note that it is possible to export individual corpora as well as groups of corpora filtered by specific conditions, such as:

- source and target languages (the platform is ready to accept not only English-Croatian corpora, but the languages can be customized),
- type of corpus (manual translation, corpus, automatically aligned, etc.),
- domain (news, manuals, tourism, song lyrics, etc.).

The export feature allows for the creation of domain-specific corpora for the needs of various natural language processing experiments and the training of machine translation systems.

The outcome of the “TMrepository” project is an extensive corpus that is meant to be applied to various natural language processing tasks. The main purpose, however, was to use it in machine translation. This purpose strongly influenced the design of the platform and the type of data it stores.

First of all, machine translation requires extensive data. State-of-the-art neural models are able to generalize over vast amounts of information to provide nearly human-quality translations. To enable this generalization, it is necessary to use significantly sized datasets for training. Hence, “TMrepository” was designed as a web platform accessible by multiple researchers at once. The collective work of many researchers allowed for the collection of data in an order of magnitude suitable for statistical and neural machine translation training.

The other aspect of the “TMrepository” that was specifically crafted for machine translation training is the organization of data by domains. Machine translation models are known to perform better when used on data coming from a single domain. This platform allows for the export of data filtered by one or more domains.

And most importantly, it was created to collect datasets for a low-resourced language pair, English-Croatian. As opposed to projects focused solely on the accumulation of parallel corpora by crawling and aligning texts from the internet, “TMrepository” also values the quality of the corpora. The result is a dataset that is potentially very interesting from the point of view of machine translation system developers.

4. Experimental Scenario

This paper presents a study on applying the concepts of crowdsourcing and gamification to a group of students with the use of “TMrepository”. The initial experiments were conducted in Poland. The students taking part in the experiment were participating in an academic course on Natural Language Processing, as part of their computer science studies.

They had completed four to six semesters of study prior to the experiment. Thus, their backgrounds covered areas such as basic algorithms and C++ programming, object-oriented programming, web applications, mathematical analysis, algebra, logic, and set theory. The students, however, had no previous training in linguistics, machine translation, or NLP. Moreover, none of them spoke Croatian, and all were native speakers of Polish. Despite certain similarities between Croatian and Polish, two Slavic languages, speakers of just one of these languages cannot fully understand the other.

During the course, the fundamentals of web scraping were covered in lectures using command-line tools such as *wget* and Python’s *urllib* module. The web crawling software framework *PyCrawler*, which can be used to crawl corpora from the web, was also presented during a lecture. After the lectures, students were asked to start an NLP project of their choice. Building translation memories for the Croatian-English language pair was one of the suggested tasks. The students that selected this assignment were told to create translation memories by “any means necessary” and by using knowledge learned during lectures.

After conducting initial experiments in Poland, additional corpora acquisition was carried out in Croatia with the help of students and experienced researchers with a focus on natural language processing. Here, the researchers were recruited for the project predominantly based on their experience using automated tools for corpus creation. This was the single mandatory skill that enabled the participants to produce valuable linguistic resources. Besides that, in summary, the profile of all contributors varied by:

- nationality—all participating students and researchers came from Poland and Croatia;
- level of language understanding—all students and researchers had at least intermediate English understanding skills, but only Croats could read and fully understand Croatian (even though the Polish and Croatian languages exhibit some similarities, they are not mutually intelligible);
- experience—from students in the early stages of their studies and graduate students in their twenties to experienced natural language processing researchers;
- gender—the distribution of women and men among the participants was nearly even;
- occupation—participants were either studying or researching the fields of information and communication sciences, computer science, linguistics, or data science with a special focus on natural language processing.

The goal of this paper was to present a platform for language resource acquisition and analyze the main characteristics of the collected data. As the work on “TMrepository” is still ongoing, the authors plan to conduct more experiments with regard to its usability, user-friendliness, and effectiveness. Furthermore, once the quality assurance phase is complete, the resulting corpus will be used to train and evaluate a group of machine translation engines for multiple translation domains.

5. Collected Corpus

Besides creating the web-based platform "TMrepository", which was designed for facilitating the collection of parallel corpora, the results of this research also include a four-million-segment English-Croatian parallel corpus. Precisely 4,091,227 translation units, which are comprised of almost 110 million words, were collected during this study (Table 2).

Table 2. Collected corpus—broken down by domains.

Domain	Segments (TUs)	Words
General	1,091,756	31,595,050
Technical	668,991	10,693,764
Tourism	636,896	22,388,296
Manuals	524,186	7,002,834
Books	491,425	15,557,436
News	364,278	15,103,005
Web	139,488	5,345,973
Song lyrics	75,465	720,750
Legal—law	62,133	1,038,258
Film subtitles	36,539	389,094
Literature—creative	70	1976
Total	4,091,227	109,836,436

The most common sources for parallel corpora include the following:

- Croatian-English and English-Croatian parallel corpora, such as SETIMES or TED Talks,
- technical documentation for various products,
- tourism websites,
- manuals,
- song lyrics,
- legal documents.

All the resources collected on "TMrepository" originate from publicly available data and open-source materials that were gathered by researchers who were willingly participating in this open-source project. Their original work is therefore not copyrighted and does not violate laws or regulations. In addition, collecting data from the internet is a standard procedure in web crawling.

The differences in translation memory size, the diversity of domains and domain independence, the variations in language register and style, and the ability to update the resources that have been collected are the main characteristics of the acquired parallel data. As initially expected by the authors, the most represented domain is "General", since accessing this type of corpora is easy. It contains non-specific data that covers a wide range of generic topics (e.g., from news), and this is usually a good starting point for building general-purpose machine translation systems.

The following domains are similar in size: "Technical", "Tourism", "Manuals", and "Books", due to the availability of bilingual resources on the internet. The domain "Technical" consists mostly of standards and guidelines related to the ICT industry, while "Tourism" was predominantly collected from tourist web sites. The domain "Manuals" contains multilingual manuals for devices and home appliances, whereas data from the domain "Books" was mainly collected from open libraries and e-book platforms.

Next in line is the "News" domain, which was primarily collected by scraping online newspapers and internet portals. The least represented domains were "Song lyrics" and "Legal—law" (again, similar in size), followed by "Film subtitles" and "Literature—creative".

Each collected domain can further be used for conducting task-oriented research, e.g., for building domain-specific machine translation systems, CAT tools, topic detection, terminology extraction, data analyses, NLP, etc.

6. Conclusions and Future Research

The main goal of this paper was to present a four-million-segment (almost 110 million words) English-Croatian corpus. It was built using a newly created web-based platform that works with other languages as well. In order to investigate the viability of a realistic implementation of software for collecting, storing, and organizing linguistic data, the authors examined the significance and various use-cases of parallel corpora, particularly for the purpose of machine translation. The platform integrates and combines the concepts of crowdsourcing and gamification, making it appropriate for both practical use by large audiences and for educational purposes. This is especially true for students that deal with natural language processing, machine translation, linguistics, CAT tools, language resources, etc. The platform has a user-friendly interface, is free to use, and is available to all users regardless of language. It facilitates international collaboration since the number of domains and language pairs can be expanded arbitrarily.

However, the platform is an ongoing project, so for future research, the authors plan to include additional functionalities, and to motivate new potential users to actively contribute to the rise of the Croatian-English corpus, especially for domains that are hardly available. In addition, the authors intend to incorporate more gamification elements into the web platform, such as challenges, avatars, levels, points, etc., to maximize the positive aspects of this methodology, to make the platform more attractive, and to encourage more users to participate by turning a boring task into a fun and entertaining one.

The corpus collected on "TMrepository" is due to be made public under an appropriate open-source license in the near future. The quality assurance process is still in progress, but once it is finished, it will be possible to release resources of optimal quality. This is all being implemented in order to create a place for preserving the resources of endangered languages.

Author Contributions: Conceptualization, R.J., S.S. and I.D.; methodology, R.J., S.S. and I.D.; software, R.J.; validation, R.J., S.S. and I.D.; formal analysis, R.J., S.S. and I.D.; investigation, R.J., S.S. and I.D.; resources, R.J., S.S. and I.D.; data curation, R.J.; writing—original draft preparation, R.J., S.S. and I.D.; writing—review and editing, I.D.; visualization, R.J., S.S. and I.D.; supervision, I.D.; funding acquisition, I.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Online, 5–10 July 2020; pp. 6282–6293. [CrossRef]
2. Haddow, B.; Bawden, R.; Barone, A.V.M.; Helcl, J.; Birch, A. Survey of Low-Resource Machine Translation. *Comput. Linguist.* **2022**, *48*, 673–732. [CrossRef]
3. Hedderich, M.A.; Lange, L.; Adel, H.; Strötgen, J.; Klakow, D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Mexico City, Mexico, 6–11 June 2021; pp. 2545–2568. [CrossRef]
4. Volk, M. Parallel Corpora, Terminology Extraction and Machine Translation. In Proceedings of the 16. DTT-Symposion. Terminologie und Text(e), Mannheim, Germany, 22–24 March 2018; pp. 3–14. [CrossRef]

5. Jaworski, R.; Seljan, S.; Dunđer, I. Usability Analysis of the Concordia Tool Applying Novel Concordance Searching. In Proceedings of the International Conference on Information Technology & Systems (ICITS 2021), Libertad, Ecuador, 4–6 February 2021; pp. 128–138. [CrossRef]
6. Macken, L.; Prou, D.; Tezcan, A. Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process. *Informatics* **2020**, *7*, 12. [CrossRef]
7. Seljan, S.; Erdelja, N.Š.; Kučiš, V.; Dunđer, I.; Bach, M.P. Quality Assurance in Computer-Assisted Translation in Business Environments. In *Natural Language Processing for Global and Local Business*; Pinarbasi, F., Nurdan Taskiran, M., Eds.; IGI Global Hershey: Hershey, PA, USA, 2021; pp. 242–270. [CrossRef]
8. Eo, S.; Park, C.; Moon, H.; Seo, J.; Lim, H. Comparative Analysis of Current Approaches to Quality Estimation for Neural Machine Translation. *Appl. Sci.* **2021**, *11*, 6584. [CrossRef]
9. Elmakias, I.; Vilenchik, D. An Oblivious Approach to Machine Translation Quality Estimation. *Mathematics* **2021**, *9*, 2090. [CrossRef]
10. Wang, Y.; Li, X.; Yang, Y.; Anwar, A.; Dong, R. Hybrid System Combination Framework for Uyghur–Chinese Machine Translation. *Information* **2021**, *12*, 98. [CrossRef]
11. Seljan, S.; Dunđer, I. Automatic quality evaluation of machine-translated output in sociological-philosophical-spiritual domain. In Proceedings of the Iberian Conference on Information Systems and Technologies (CISTI 2015), Aveiro, Portugal, 17–20 June 2015; pp. 1–4. [CrossRef]
12. Jaworski, R.; Seljan, S.; Dunđer, I. Towards educating and motivating the crowd—A crowdsourcing platform for harvesting the fruits of NLP students' labour. In Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2017), Poznań, Poland, 17–19 November 2017; pp. 332–336.
13. Kučiš, V.; Seljan, S. The role of online translation tools in language education. *Babel* **2014**, *60*, 303–324. [CrossRef]
14. Gašpar, A.; Seljan, S.; Kučiš, V. Measuring Terminology Consistency in Translated Corpora: Implementation of the Herfindahl-Hirshman Index. *Information* **2022**, *13*, 43. [CrossRef]
15. Béchara, H.; Orăsan, C.; Parra Escartín, C.; Zampieri, M.; Lowe, W. The Role of Machine Translation Quality Estimation in the Post-Editing Workflow. *Informatics* **2021**, *8*, 61. [CrossRef]
16. Han, B. Translation, from Pen-and-Paper to Computer-Assisted Tools (CAT Tools) and Machine Translation (MT). *Proceedings* **2020**, *63*, 56. [CrossRef]
17. Wang, R.; Tan, X.; Luo, R.; Qin, T.; Liu, T.-Y. A Survey on Low-Resource Neural Machine Translation. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Online, 19–26 August 2021; pp. 4636–4643. [CrossRef]
18. Ngo, T.V.; Nguyen, P.-T.; Ha, T.-L.; Dinh, K.-Q.; Nguyen, L.-M. Improving Multilingual Neural Machine Translation For Low-Resource Languages: French, English—Vietnamese. In Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages, Suzhou, China, 4 December 2020; pp. 55–61. [CrossRef]
19. Ranathunga, S.; Lee, E.-S.A.; Prifti Skenduli, M.; Shekhar, R.; Alam, M.; Kaur, R. Neural Machine Translation for Low-Resource Languages: A Survey. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
20. Koehn, P.; Och, F.J.; Marcu, D. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03), Edmonton, AB, Canada, 27 May–1 June 2003; Volume 1, pp. 127–133. [CrossRef]
21. Kamath, U.; Liu, J.; Whitaker, J. *Deep Learning for NLP and Speech Recognition*; Springer: Berlin/Heidelberg, Germany, 2019. [CrossRef]
22. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015; pp. 1–15. [CrossRef]
23. Koehn, P. *Statistical Machine Translation*; Cambridge University Press: New York, NY, USA, 2010. [CrossRef]
24. Dong, D.; Wu, H.; He, W.; Yu, D.; Wang, H. Multi-Task Learning for Multiple Language Translation. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Beijing, China, 27–31 July 2015; Volume 1, pp. 1723–1732. [CrossRef]
25. Gehring, J.; Auli, M.; Grangier, D.; Dauphin, Y. A Convolutional Encoder Model for Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 123–135. [CrossRef]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010. [CrossRef]
27. Singh, T.D. Building Parallel Corpora for SMT System: A Case Study of English-Manipuri. *Int. J. Comput. Appl.* **2012**, *52*, 47–51. [CrossRef]
28. Dunđer, I. Statistical Machine Translation System and Computational Domain Adaptation (Sustav za Statističko Strojno Prevođenje i Računalna Adaptacija Domene). Ph.D. Thesis, University of Zagreb, Zagreb, Croatia, 2015.
29. Parida, S.; Bojar, O.; Dash, S.R. OdiEnCorp: Odia–English and Odia-Only Corpus for Machine Translation. In Proceedings of the Third International Conference on Smart Computing and Informatics (SCI 2018-19), Bhubaneswar, India, 21–22 December 2018; Volume 1, pp. 495–504. [CrossRef]

30. Ambati, V.; Vogel, S. Can Crowds Build Parallel Corpora for Machine Translation Systems? In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10), Los Angeles, NY, USA, 6 June 2010; pp. 62–65.
31. Abdurakhmonova, N. Linguistic Issues of Creating Parallel Corpora for Uzbek Multilingual Machine Translation System. *BuxDU Ilmiy Axborotnomasi* **2020**, *6*, 60–68.
32. Doğru, G.; Martín-Mor, A.; Aguilar-Amat, A. Parallel Corpora Preparation for Machine Translation of Low-Resource Languages: Turkish to English Cardiology Corpora. In Proceedings of the LREC 2018 Workshop 'MultilingualBIO: Multilingual Biomedical Text Processing', Miyazaki, Japan, 7–12 May 2018; pp. 12–15.
33. Shearing, S.; Kirov, C.; Khayrallah, H.; Yarowsky, D. Improving Low Resource Machine Translation using Morphological Glosses. In Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, Boston, MA, USA, 17–21 March 2018; Volume 1, pp. 132–139.
34. Forcada, M.L. Free/Open-Source Machine Translation for the Low-Resource Languages of Spain. In Proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021), Zaragoza, Spain, 1–4 September 2021. [CrossRef]
35. Chu, C.; Wang, R. A Survey of Domain Adaptation for Neural Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, NM, USA, 20–26 August 2018; pp. 1304–1319. [CrossRef]
36. Dabre, R.; Chu, C.; Kunchukuttan, A. A survey of multilingual neural machine translation. *ACM Comput. Surv.* **2020**, *53*, 99. [CrossRef]
37. Maruf, S.; Saleh, F.; Haffari, G. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Comput. Surv.* **2021**, *54*, 45. [CrossRef]
38. Kuwanto, G.; Akyürek, A.F.; Tourni, I.C.; Li, S.; Jones, A.G.; Wijaya, D. Low-Resource Machine Translation Training Curriculum Fit for Low-Resource Languages. *arXiv* **2021**, arXiv:2103.13272. [cs.CL], Computation and Language. [CrossRef]
39. Sen, S.; Hasanuzzaman, M.; Ekbal, A.; Bhattacharyya, P.; Way, A. Neural machine translation of low-resource languages using SMT phrase pair injection. *Nat. Lang. Eng.* **2020**, *27*, 271–292. [CrossRef]
40. Beloucif, M.; Gonzalez, A.V.; Bollmann, M.; Søgaard, A. Naive Regularizers for Low-Resource Neural Machine Translation. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; pp. 102–111. [CrossRef]
41. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. In Proceedings of the First Workshop on Neural Machine Translation, Vancouver, BC, Canada, 3–4 July 2017; pp. 28–39. [CrossRef]
42. Seljan, S.; Dunđer, I.; Pavlovski, M. Human Quality Evaluation of Machine-Translated Poetry. In Proceedings of the International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 28 September–2 October 2020; pp. 1040–1045. [CrossRef]
43. Lambebo, A.; Woldeyohannis, M.; Yigezu, M. A Parallel Corpora for bi-directional Neural Machine Translation for Low Resourced Ethiopian Languages. In Proceedings of the 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 22–24 November 2021; pp. 71–76. [CrossRef]
44. Zhang, J.; Tian, Y.; Mao, J.; Han, M.; Wen, F.; Guo, C.; Gao, Z.; Matsumoto, T. WCC-JC 2.0: A Web-Crawled and Manually Aligned Parallel Corpus for Japanese-Chinese Neural Machine Translation. *Electronics* **2023**, *12*, 1140. [CrossRef]
45. Ha, T.-L.; Niehues, J.; Waibel, A. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In Proceedings of the 13th International Conference on Spoken Language Translation (IWSLT 2016), Seattle, DC, USA, 8–9 December 2016. [CrossRef]
46. Lakew, S.M.; Cettolo, M.; Federico, M. A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 641–652. [CrossRef]
47. Tan, X.; Ren, Y.; He, D.; Qin, T.; Zhao, Z.; Liu, T.-Y. Multilingual Neural Machine Translation with Knowledge Distillation. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 6–9 May 2019; pp. 1–15. [CrossRef]
48. Aji, A.F.; Bogoychev, N.; Heafield, K.; Sennrich, R. In Neural Machine Translation, What Does Transfer Learning Transfer? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), Online, 5–10 July 2020; pp. 7701–7710. [CrossRef]
49. Kim, Y.; Gao, Y.; Ney, H. Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, 28 July–2 August 2019; pp. 1246–1257. [CrossRef]
50. Dabre, R.; Nakagawa, T.; Kazawa, H. An Empirical Study of Language Relatedness for Transfer Learning in Neural Machine Translation. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC 2017), Manila, Philippines, 16–18 November 2017; pp. 282–286.
51. Zoph, B.; Yuret, D.; May, J.; Knight, K. Transfer Learning for Low-Resource Neural Machine Translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), Austin, TX, USA, 1–5 November 2016; pp. 1568–1575. [CrossRef]

52. Wang, W.; Zhang, Z.; Du, Y.; Chen, B.; Xie, J.; Luo, W. Rethinking Zero-shot Neural Machine Translation: From a Perspective of Latent Variables. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 4321–4327. [CrossRef]
53. Currey, A.; Heafield, K. Zero-Resource Neural Machine Translation with Monolingual Pivot Data. In Proceedings of the 3rd Workshop on Neural Generation and Translation (NGT 2019), Hong Kong, China, 3–7 November 2019; pp. 99–107. [CrossRef]
54. O'Brien, S. Collaborative translation. In *Handbook of Translation Studies*; Gambier, Y., van Doorslaer, L., Eds.; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2011; Volume 2, pp. 17–20. [CrossRef]
55. Howe, J. The Rise of Crowdsourcing. *Wired Mag.* **2006**, *14*. Available online: http://www.wired.com/wired/archive/14.06/crowds.html (accessed on 11 February 2023).
56. Howe, J. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, 1st ed.; Crown Publishing Group: New York, NY, USA, 2008.
57. Quinn, A.J.; Bederson, B.B. Human Computation: A Survey and Taxonomy of a Growing Field. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11), New York, NY, USA, 7–12 May 2011; pp. 1403–1412. [CrossRef]
58. Sabou, M.; Bontcheva, K.; Derczynski, L.; Scharl, A. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; pp. 859–866.
59. Li, H.; Shen, H.; Xu, S.; Zhang, C. Visualizing NLP annotations for Crowdsourcing. *arXiv* **2015**, arXiv:1508.06044. [cs.CL], Computation and Language. [CrossRef]
60. Munro, R.; Gunasekara, L.; Nevins, S.; Polepeddi, L.; Rosen, E. Tracking Epidemics with Natural Language Processing and Crowdsourcing. In Proceedings of the AAAI Spring Symposium—Wisdom of the Crowd (AAAI 2012), Palo Alto, CA, USA, 26–28 March 2012.
61. Sabou, M.; Bontcheva, K.; Scharl, A. Crowdsourcing Research Opportunities: Lessons from Natural Language Processing. In Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '12), Graz, Austria, 5–7 September 2012; pp. 1–8. [CrossRef]
62. Vamshi, A.; Vogel, S.; Carbonell, J. Collaborative Workflow for Crowdsourcing Translation. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12), Seattle, DC, USA, 11–15 February 2012; pp. 1191–1194.
63. Zaidan, O.F.; Callison-Burch, C. Crowdsourcing Translation: Professional Quality from Non-professionals. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11), Portland, OR, USA, 19–24 June 2011; Volume 1, pp. 1220–1229.
64. Muntés-Mulero, V.; Paladini, P.; Solé, M.; Manzoor, J. Multiplying the Potential of Crowdsourcing with Machine Translation. In Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Commercial MT User Program (AMTA 2012), San Diego, CA, USA, 28 October–1 November 2012.
65. Muegge, U. Teaching computer-assisted translation in the 21st century. In *TransÜD. Arbeiten zur Theorie und Praxis des Übersetzens und Dolmetschens (Alles Hängt Mit Allem Zusammen: Translatologische Interdependenzen. Festschrift für Peter A. Schmitt)*; Ende, A.-K., Herold, S., Weilandt, A., Eds.; Frank & Timme: Berlin, Germany, 2013; Volume 59.
66. Canovas, M.; Samson, R. Open source software in translator training. *Rev. Tradumàtica* **2011**, *9*, 6–56. [CrossRef]
67. Robson, K.; Plangger, K.; Kietzmann, J.H.; McCarthy, I.; Pitt, L. Game on: Engaging customers and employees through gamification. *Bus. Horiz.* **2016**, *59*, 29–36. [CrossRef]
68. Morschheuser, B.; Werder, K.; Hamari, J.; Abe, J. How to gamify? Development of a method for gamification. In Proceedings of the 50th Annual Hawaii International Conference on System Sciences (HICSS), Hawaii, HI, USA, 4–7 January 2017; Volume 50, pp. 1298–1307. [CrossRef]
69. Abdelali, A.; Durrani, N.; Guzmán, F. iAppraise: A Manual Machine Translation Evaluation Environment Supporting Eye-tracking. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (NAACL 2016), San Diego, CA, USA, 13–15 June 2016; pp. 17–21. [CrossRef]
70. Graliński, F.; Jaworski, R.; Borchmann, Ł.; Wierzchon, P. Gonito.net—Open platform for research competition, cooperation and reproducibility. In Proceedings of the 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language, Portorož, Slovenia, 28 May 2016; pp. 13–20.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.