data

MDPI

Data Descriptor

A Comprehensive Dataset of Spelling Errors and Users' Corrections in Croatian Language

Gordan Gledec [1,*], Marko Horvat [1], Miljenko Mikuc [2] and Bruno Blašković [3]

[1] Department of Applied Computing, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia; marko.horvat3@fer.hr
[2] Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia; miljenko.mikuc@fer.hr
[3] Department of Electrical Engineering Fundamentals and Measurements, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia; bruno.blaskovic@fer.hr
* Correspondence: gordan.gledec@fer.hr

Abstract: This paper presents a unique and extensive dataset containing over 33 million entries with pairs in the form "spelling error → correction" from ispravi.me, the most popular Croatian online spellchecking service, collected since 2008. The dataset, compiled from the contribution of nearly 900,000 users, is a valuable resource for researchers and developers in the field of natural language processing (NLP), improving spellcheck accuracy, and language learning applications. The dataset may be used to accomplish several goals: (1) improving spellchecking accuracy by incorporating common user corrections and reducing false positives and negatives; (2) helping language learners identify common errors and learn correct spelling through targeted feedback; (3) analyzing data trends and patterns to uncover the most common spelling errors and their underlying causes; (4) identifying and evaluating factors that influence typing input; (5) improving NLP applications such as text recognition and machine translation. Tasks specific to the Croatian language include the creation of a letter-level confusion matrix and the refinement of word suggestions based on historical usage of the service. This comprehensive dataset provides researchers and practitioners with a wealth of information, opening the path for advancements in spellchecking, language learning, and NLP applications in the Croatian language.

Dataset: https://github.com/Ispravi-Me/Dataset-of-Misspelings-and-Corrections

Dataset License: CC BY-NC-SA 4.0

Keywords: spellchecker; n-grams; natural language processing; Croatian language; user corrections dataset; common error analysis

check for updates

Citation: Gledec, G.; Horvat, M.; Mikuc, M.; Blašković, B. A Comprehensive Dataset of Spelling Errors and Users' Corrections in Croatian Language. *Data* **2023**, *8*, 89. https://doi.org/10.3390/data8050089

Academic Editor: Joaquín Torres-Sospedra

Received: 14 April 2023
Revised: 28 April 2023
Accepted: 10 May 2023
Published: 12 May 2023

1. Introduction

The increasing prevalence of digital communication has highlighted the importance of accurate spelling and grammar in written text. Spellchecking services play a crucial role in ensuring the quality and readability of digital content, aiding both native speakers and language learners in producing error-free text [1]. This is particularly crucial for languages such as Croatian, which exhibit complex morphological and orthographic rules [2].

In the context of natural language processing (NLP) technologies, it is essential to acknowledge the distinct differences between individual languages. For example, when considering the Croatian language, features specifically developed for the Chinese (for example, [3]) would be irrelevant due to the language's distinct characteristics including its Slavic origin, Latin script, and phonetic orthography. Consequently, error correction algorithms and techniques applied to Chinese cannot be directly used for Croatian without significant adaptations to account for the fundamental differences between the two languages.

Currently, the availability of language technologies for Croatian is limited, lacking support in the areas of machine translation, speech processing, and text analysis, while offering only partial support in speech and text resources [4]. The detailed state of language technologies for the Croatian language was described in [5], and since then, several contributions have been made in recent years. The Natural Language Processing group at the University of Zagreb's Faculty of Humanities and Social Sciences created a high-quality web corpus, MaCoCu-Hr, for linguistic studies, training language models, and other language applications [6]. Additionally, researchers at the University of Zagreb Faculty of Electrical Engineering and Computing have developed an n-gram language model [7,8] and a contextual spellchecking model based on the n-gram system [9]. Finally, a group from the University of Rijeka examined sentiments in Twitter tweets during the COVID-19 pandemic and developed their own model by utilizing BERT tools and a linguistic corpus [10]. Our comprehensive dataset of spelling errors and user corrections presented in the paper aims to further advance research in language technologies for the Croatian language.

Typos are the most common errors, which occur when people accidentally insert, omit, replace, or transpose letters in a word [11]. A common cause of spelling errors is a lack of competence related to the language's orthography rules or medical conditions such as dyslexia, dysgraphia, or other cognitive disorders [12]. Different languages, however, are prone to different types of spelling errors [13]. Croatian, like most Slavic languages, has letters with diacritics or accents that users omit when writing, preferring simpler variations for speed without sacrificing readability. Such letters, which lack diacritics and accents, are easily accessible on the keyboard and do not necessitate multiple keystrokes. Although on most occasions such replacements result in an error word, they sometimes form another correct word and result in a real-word error (e.g., "što → sto", "pošto → posto", etc.). This problem was dealt with by Šantić et al., who developed a system for automatic diacritics restoration in the Croatian language [14].

Ispravi.me (English Correct.me) [15] is a pioneering Croatian spellchecking service that was introduced in the early 1990s under the name of Hascheck (as an abbreviation of the Hrvatski akademski spelling checker—Croatian Academic Spelling Checker). Starting from a corpus of 100,000 Croatian words. It has since grown into a widely used online platform with more than 7000 daily user sessions as of March 2023. In fact, ispravi.me is currently the most popular spellchecking service for the Croatian language and is frequently used by major domestic news services. Because of the similarities among the South Slavic languages, it is also used in other countries bordering Croatia.

Since the ispravi.me service learns new words from user text, the growing crowd-sourced database includes modern words, slang, abbreviations, named entities, and many other spelling and grammar artifacts. The service not only marks unknown words, but also allows users to select the correct word from a dropdown menu, logging the "error word → correct word" pair on the server side in this process. By logging what users select from the drop-down list of possible substitutions, it is possible to deduce which intended word actually ended up as the wrong or mistyped word. This feature has yielded a rich dataset of over 33 million entries including five million typos and 1.5 million corrections generated by nearly 900,000 users over the last 15 years.

When researching spellchecking in the 1980s, Roger Mitton was questioned about why he did not compile a sizable database of misspellings to map them onto their target terms, which would have sped up the procedure [16]. Given the creative possibilities of misspellings, Mitton contended that compiling such a collection would be a monumental undertaking and an uncontrollably large database. In recent years, Mitton published a corpora of misspellings available for download at [17], with common misspellings from four different sources. In their paper on using web data for spellchecking in the Kazakh language, Toleu et al., also argued that constructing this type of manually annotated corpora is time consuming and often expensive [18].

Thanks to the Internet and our spellchecking application, a database comparable to Mitton's envisioned database now exists for Croatian, exactly as Mitton had envisioned thirty years ago: "So perhaps, when you make a spelling error and the correct spelling pops into your computer, it may be that you will be benefiting not so much from the efforts of good spellers who have gone before you, patiently creating dictionaries of correct spellings, but from the efforts of bad ones, misspelling the same word in a thousand different ways" [16].

This paper presents and provides a thorough examination of the ispravi.me dataset, providing useful insights into common spelling errors, user corrections, and trends in the Croatian language. We investigate how this dataset could be used to improve the accuracy of spellcheckers, assist language learners, identify trends in spelling mistakes, and improve natural language processing applications such as text recognition and voice recognition software.

The rest of this paper is organized as follows. Section 2 describes the dataset, its attributes and format. Section 3 explains the methodology employed in collecting and processing the data. Section 4 discusses the potential applications of the dataset in various domains including spellchecker accuracy improvement, language learning, and natural language processing. This section also delves into specific tasks related to the Croatian language such as the development of a letter-level confusion matrix and improved word suggestions. Finally, Section 5 concludes the paper and outlines future research directions.

2. Data Description

The data contained in our tab-delimited dataset is described in Table 1. Each attribute has a specific function in characterizing user-generated spelling errors and their corrections. The dataset is encoded in UTF-8 charset.

Table 1. Description of the attributes in the dataset.

Attribute	Description
date	The date in YYYY-MM-DD format when the entry is logged.
UserID	Before 27 November 2016: 0 After 27 November 2016: universally unique identifier (UUID) of the user, represented as 32 hexadecimal digits, using uppercase, displayed in five groups separated by hyphens, in the form 8-4-4-4-12.
error_word	The original unknown word that user later replaced with one of the suggested words.
correct_word	The word the user chose from the dropdown menu of suggested words to replace the mistyped word.
edit_distance	Damerau–Levenshtein edit distance between the error word and the correct word—possible values are either 1 or 2.

The "date" attribute stores the date in the YYYY-MM-DD format. The 'UserID' attribute uniquely distinguishes individual users by using a Universally Unique Identifier (UUID) consisting of 32 uppercase hexadecimal digits separated by hyphens into five groups (8-4-4-4-12) [19]. This maintains anonymity while enabling user-specific analysis. The UUID was introduced on 27 November 2016. In our database, the value of this attribute prior to that date is set to 0.

The "error_word" records the original unknown or mistyped word that the user later replaces with a suggested alternative.

The "correct_word" attribute represents the word selected by the user from the suggestions in the drop-down menu that reflects the intended meaning or correct spelling.

For example, in a recorded pair “zamijena → zamjena”, the error word is “zamijena”, and the correct word is “zamjena”. Sometimes, an error word is corrected by multiple different words, for example, “zamijena” can also be corrected with “zamijeni”, “zamijene”, “zabijena”, “zavijena”, and “zamiješa” (in English, “[you] replace”, “[they] replace”, “nailed down”, “wrapped”). Of course, a single correct word may be the result of many different error words.

Finally, the “edit_distance” attribute calculates the Damerau–Levenshtein edit distance [11] between the error and correct words. The Damerau–Levenshtein distance is a metric for measuring the edit distance between two strings as the minimum number of insertions, deletions, substitutions, or transpositions of a single character required to transform one string into another [12]. The values of this attribute can be either 1 or 2. This attribute provides useful information about the nature and complexity of spelling errors encountered by users.

It should be noted that the privacy and anonymity of users of the ispravi.me spelling service is guaranteed. Although we used UUIDs, no private information was collected from our users. In addition, users are not currently required to register or log in to use the service. UUIDs were only used to distinguish between users for research purposes. In addition, incoming texts were deleted from our servers once text processing was complete. This is clearly stated in our privacy policy, which is available online.

However, since UUIDs are transferred via cookies, one needs to keep in mind their shortcomings and where the limits of their use are:

- A user can have multiple cookies, one for each browser used;
- Users can disable or delete cookies in their browsers, so that a new cookie is set for each request from the same user;
- Multiple users can have the same cookie if some kind of intermediate application is used.

Despite these shortcomings in the use of cookies, they may be helpful in the user-specific analysis of spelling errors.

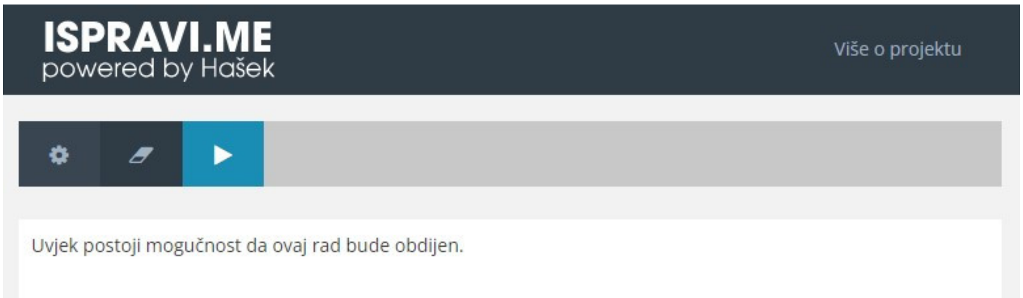
3. Methods

3.1. Data Collection

The presented dataset was collected using the ispravi.me spellchecking service. The service identifies unknown words and grammar errors and enables users to choose the correct word from a dropdown menu, logging the “error word → correct word” pair. By tracking user selections, it deduces the intended words that were mistyped or incorrect. As mentioned earlier, the ispravi.me service is currently available online at https://ispravi.me/ (English correct.me) and according to our collected server statistics and Google Analytics data, as of March 2023, it serves more than 7000 user sessions per day. It has a stable user base with more than 80% returning visitors. From May 2007 up to March 2023, the service processed 28 million texts, which form a corpus of 7.5 gigatokens (Gtokens). Overall, the service was accessed by 1.5 million users in 8.6 million sessions.

The data in the presented dataset were collected following the typical usage scenario for the spellcheck service:

1. When a user wants to check the spelling of a text, they write (or pastes) the text into the online form (Figure 1).



Uvjek postoji mogučnost da ovaj rad bude obdijen.

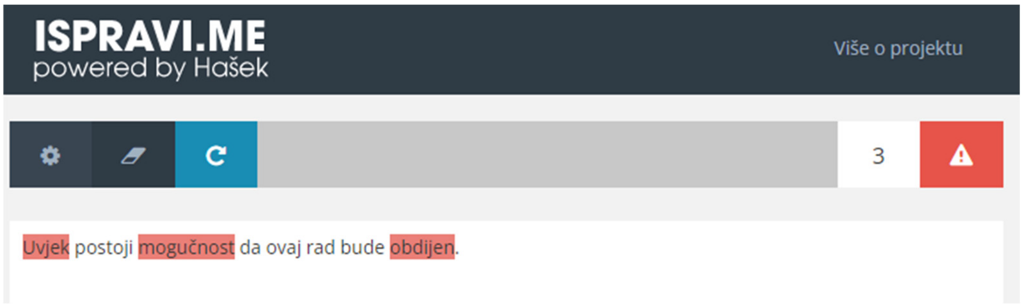
Figure 1. Service user interface—the user enters the text for spellchecking.

2. The entered text is stripped of all markup tags, formatting, and most punctuation, sent to the server via an Ajax call, and processed by the spellchecking application [20]. The inner workings of the spellchecker are described in [2,8,20,21].
3. The result is delivered to the client in JSON format, parsed by client-side JavaScript, and displayed in the browser.

An excerpt from the response for the misspelled text "Uvjek postoji mogučnost da ovaj rad bude obdijen" (should be "Uvijek postoji mogućnost da ovaj rad bude odbijen", Engl. "There is always a possibility that this paper will be rejected") in JSON format is:

```
{
"response" : {
"errors" : 4,
"error" : [
{
"position" : [0],
"length" : 5,
"suspicious" : "Uvjek",
"suggestions" : ["Uvijek","Uv'jek","Usjek","Uvjet"],
"class" : "minor",
"occurrences" : 1
},
{
"position" : [14],
"length" : 9,
"suspicious" : "mogučnost",
"suggestions" : ["mogućnost"],
"class" : "major",
"occurrences" : 1,
},
{
"position" : [41],
"length" : 7,
"suspicious" : "obdijen",
"suggestions" : ["odbijen","obijen","obvijen"],
"class" : "moderate",
"occurrences" : 1,
},
]
}
}
```

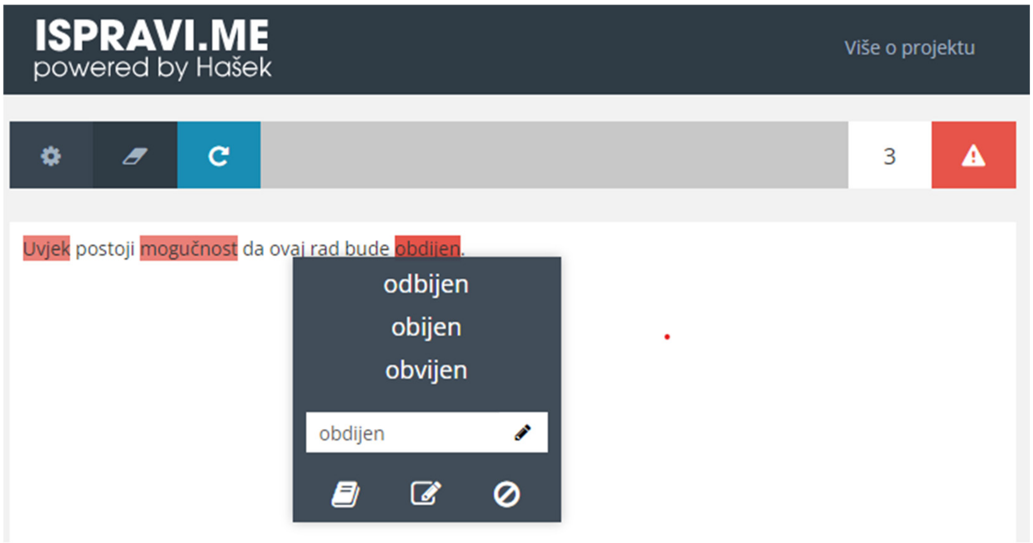
4. Upon receiving the formatted JSON response, the client interface displays the entered text to the user, with suspicious words and phrases clearly marked on the screen (Figure 2).



Uvjek postoji mogučnost da ovaj rad bude obdijen.

Figure 2. Service user interface—the text is checked and errors are highlighted.

5. By clicking on the highlighted word, a popup appears with suggested corrections for a possibly misspelled word (Figure 3): "odbijen" (rejected), "obijen" (broken open), "obvijen" (enveloped, enclosed).



Uvjek postoji mogučnost da ovaj rad bude obdijen.

Figure 3. Service user interface—the possible corrections are presented to the user.

6. The user can then click and select the correct word from the candidates, or enter the correct word into the input field in the pop-up menu if the service has not responded with the correct suggestion.
7. Once the selection is made, the user's text is updated with the corrected word and the pair "error word → correct word" is sent to the server via a dedicated CGI script, which logs it and was later used together with the metadata (e.g., environment and query variables) for our research. The date stamp of the request is also logged. The output sent to the server for the example from the previous page has the following format:

```
Uvjek -> Uvijek
mogučnost -> mogućnost
obdijen -> odbijen
```

The logs are currently stored in the server's file system as flat comma-separated (CSV) Unicode text files.

3.2. Creating the Dataset

To create or update our dataset, we parsed all log files and extracted all entries in the form "error word → correct word" along with the metadata (date, UserID). We then computed the Damerau–Levenshtein edit distance (our measure of choice) between the incorrect word and the correct word, and filtered out only those pairs where the distance was 1 or 2 (since users can enter any word as a replacement word).

3.3. Size of the Dataset

Using the described method, we collected a total of 33,382,330 entries of the form "error word → correct word" between December 2008 and March 2023—93% of which were of edit distance 1 and 7% of edit distance 2. In this huge dataset, we identified 5,584,226 unique "error word → correct word" pairs. In total, 5,296,266 unique words were misspelled, which we corrected to a total of 1,530,329 words.

3.4. Dataset Distribution

The dataset is distributed as a gzipped tab-separated Unicode text file via the GitHub service. Each file contains data for one year, and the file name is in the format "ispravime-YEAR.gz". The individual file sizes are given in Table 2. The rows are sorted by date (first column, YYYY-MM-DD) in ascending order. The dataset is updated every year when new files are added to the dataset.

Table 2. The number of records and the file sizes of the individual dataset files.

Year	No. of Records	File Size (bytes)
2008 *	2008	68,703
2009	85,906	2,917,640
2010	188,994	6,434,960
2011	315,821	10,748,864
2012	563,572	19,252,554
2013	639,414	21,940,712
2014	703,373	24,218,505
2015	794,094	27,337,889
2016	1,002,547	40,825,022
2017	2,956,906	206,155,833
2018	3,969,900	276,579,152
2019	4,565,391	318,131,677
2020	5,645,739	393,447,028
2021	5,524,501	385,070,748
2022	5,277,407	367,752,536
2023 **	1,146,757	80,006,348
Total	33,382,330	2,196,045,185

* Incomplete year; data starts on 6 December 2008; ** Incomplete year; data ends on 16 March 2023.

Figure 4 shows the growth of the database from 2009 to 2022. The rebranding of the service from Hascheck to ispravi.me, new features, new user interface as well as our social media promotional efforts and media coverage resulted in a significant increase in usage during 2016. This chart does not include the incomplete years of 2008 and 2023. The stronger increase in usage in 2020 was most likely due to the outbreak of the COVID-19 pandemic as more users stayed at home and used our service for business or school purposes.

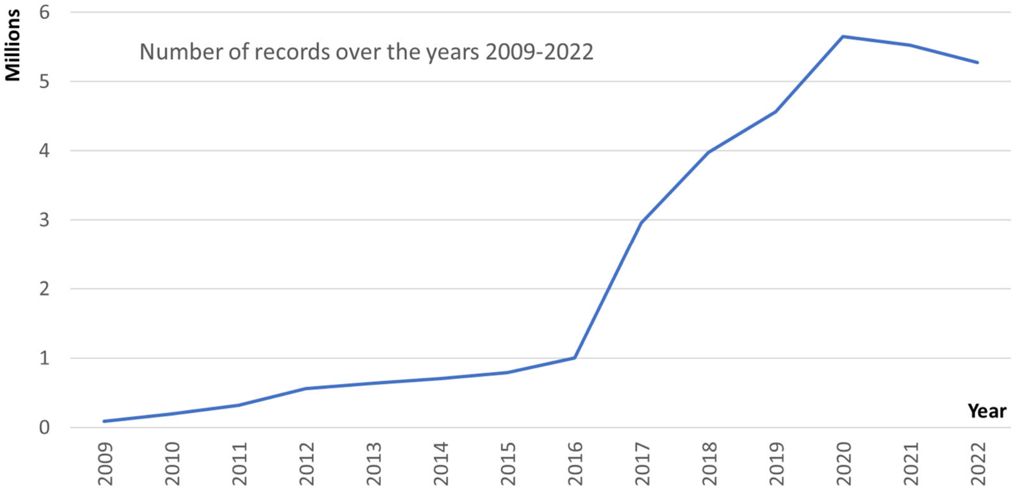


Figure 4. Growth of the ispravi.me service database from 2009 to 2022.

3.5. Noise and Garbage in the Dataset

A database of this size, without a doubt, presents its own set of challenges including noise and data inconsistencies. One factor that contributes is the ability of users to enter their own corrections when the provided suggestions do not meet their expectations, which occasionally results in the entry of another incorrect word. Additionally, some users may experiment with the interface, contributing to the data noise. Furthermore, it is important to consider that the spellchecker itself is not perfect and may occasionally produce errors.

In analyzing this dataset, we found that not all of the words that the users chose as replacement words were correct: we ran the list of "corrected" words through our spellchecker and found that 41,846 replacement words were not actually in our dictionary. This accounted for 3% of all replacement words, and since these words occurred a total of 70,000 times in our dataset, they only affected a negligible 0.2% of our dataset. Nevertheless, we kept them in our dataset.

Some of these reflect the errors in our database that resulted in accepting the error word and classifying the correct word as incorrect, with suggestions presented to the user who, not knowing the proper orthography, then actually chose the error word. For example, the word "svijetlosmeđe" was offered as a suggestion for "svjetlosmeđe" (Engl. light brown), although the latter is actually a correct word. The same applies to all colors beginning with "svijetlo-" (light). This systematic error was corrected in February 2021. Some of the correct words were learned later by the spellchecker, but until then, are present in our logs. Some words belong to the Serbian language corpus and their users chose them instead of using our suggestions (e.g., "definiše", "uopšte", "takođe", "projekat", "reči", "opštine", . . .). Some words are brand names or abbreviations.

4. Findings

The dataset presented in this paper can be used for several general purposes:

1. To improve the accuracy of spellcheck services by logging the most frequent corrections made by users, thereby reducing the number of false positives and false negatives or enabling auto-corrections of misspellings;
2. As a teaching tool for language learners to help them identify common errors and learn the correct spelling of words (e.g., a language learning app can use this database to provide instant feedback on the spelling of words entered by the user or automatically generate realistic spelling errors to test the user's knowledge);
3. For data analysis to identify trends and patterns in user errors (e.g., to identify the most frequently misspelled words or the most common types of spelling errors);

4. For evaluating factors that affect input such as the type of error (typo, orthographic, or grammatical error, . . .) and the position of the error with respect to the length of the word;
5. In natural language processing (NLP) applications to improve the accuracy of text recognition and automatic text analysis (e.g., to improve the accuracy of speech recognition software, optical character recognition (OCR) software, or to build an n-gram language model, as in the case of Polish [22,23], Czech [24,25], or Russian [26] as well as Slavic languages that also share similar research issues).

The ispravi.me dataset presented in this paper offers numerous general applications and also includes more specific tasks related exclusively to the Croatian language and our online spelling service:

1. Developing a letter-level confusion matrix for the Croatian language [27];
2. Building and maintaining an n-gram language model [8,28];
3. Improved suggestions for correct words based on past use of the service [9];
4. Facilitating the learning of spelling in the Croatian language.

Currently, we are actively developing some of these applications. At the same time, we would like to share the dataset with other researchers and encourage them to use it and create innovative applications of their own.

5. Conclusions

The presented dataset contains over 33 million entries and provides invaluable insights into the nature of spelling mistakes, user correction patterns, and Croatian language trends. The dataset includes more than five million unique spelling errors and 1.5 million unique correct replacements based on historical service usage and presents a foundation for the creation of a letter-level confusion matrix for the Croatian language. The dataset is a valuable resource for language learners, providing targeted feedback on common errors and assisting in the learning of correct spelling.

The data analysis also revealed trends and patterns in spelling errors, allowing for the identification of the most frequently misspelled words and error types. Resources from this dataset are vital for improving language learning tools and refining linguistic research. It is important to note that there are relatively few NLP resources developed specifically for the Croatian language. This makes our dataset even more important in the context of NLP advancements for Croatian as well other Southern Slavic languages.

We have identified numerous potential applications for this dataset in a variety of domains including spellchecker accuracy improvement, language learning, and natural language processing. In the broader context of natural language processing, the dataset can help improve the accuracy of text recognition and speech recognition software.

Future research directions may include further expansion and refinement of the dataset and the exploration of additional applications in language technology, education, and linguistic research. Regarding the latter, we would like to explore the development of a Croatian non-contextual word embedding model, taking inspiration from the successful implementation of a 256-dimensional vector model trained for the needs of the Kontekst.io portal in the Slovenian language [29,30]. Creating a similar model for the Croatian language could significantly improve the performance of various NLP tasks such as text classification, sentiment analysis, and machine translation. Such a contribution would be useful for both academic research and practical applications of language technology.

Author Contributions: Conceptualization, G.G. and M.H.; Methodology, G.G.; Software, G.G.; Validation, M.H., M.M. and B.B.; Formal analysis, B.B. and M.M.; Investigation, G.G.; Resources, G.G.; Data curation, G.G.; Writing—original draft preparation, G.G.; Writing—review and editing, G.G. and M.H.; Visualization, G.G.; Supervision, M.M. and B.B.; Project administration, G.G. and M.M.; Funding acquisition, G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Dembitz, Š.; Gledec, G.; Randić, M. Spellchecker. In *Wiley Encyclopedia of Computer Science and Engineering*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009.
2. Dembitz, Š.; Randić, M.; Gledec, G. Advantages of Online Spellchecking: A Croatian Example. *Softw. Pract. Exp.* **2011**, *41*, 1203–1231. [CrossRef]
3. Gou, W.; Chen, Z. Think Twice: A Post-Processing Approach for the Chinese Spelling Error Correction. *Appl. Sci.* **2021**, *11*, 5832. [CrossRef]
4. META-NET White Paper Series. Key Results and Cross-Language Comparison. Available online: http://www.meta-net.eu/whitepapers/overview (accessed on 12 April 2023).
5. Tadić, M.; Brozović-Rončević, D.; Kapetanović, A. *The Croatian Language in the Digital Age*; Rehm, G., Uszkoreit, H., Eds.; Springer: Berlin, Heidelberg, 2012; ISBN 978-3-642-30881-9.
6. Bañón, M.; Chichirau, M.; Esplà-Gomis, M.; Forcada, M.L.; Galiano-Jiménez, A.; García-Romero, C.; Kuzman, T.; Ljubešić, N.; van Noord, R.; Pla Sempere, L. Croatian Web Corpus MaCoCu-hr 2.0. Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. 2023. Available online: http://hdl.handle.net/11356/1806 (accessed on 12 April 2023).
7. Šoić, R.; Vuković, M. N-Gram Based Croatian Language Network: Application in a Smart Environment. *J. Commun. Softw. Syst.* **2022**, *18*, 63–71. [CrossRef]
8. Gledec, G.; Šoić, R.; Dembitz, Š. Dynamic N-Gram System Based on an Online Croatian Spellchecking Service. *IEEE Access* **2019**, *7*, 149988–149995. [CrossRef]
9. Srdić, I.; Gledec, G. Contextual Spellchecking Based on N-Grams. In Proceedings of the 28th Central European Conference on Information and Intelligent Systems; Faculty of Organization and Informatics, Varaždin, Croatia, 27–29 September 2017; pp. 29–33.
10. Babić, K.; Petrović, M.; Beliga, S.; Martinčić-Ipšić, S.; Matešić, M.; Meštrović, A. Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-CseBERT Model. *Appl. Sci.* **2021**, *11*, 10442. [CrossRef]
11. Damerau, F.J. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* **1964**, *7*, 171–176. [CrossRef]
12. Mitton, R. Fifty Years of Spellchecking. *Writ. Syst. Res.* **2010**, *2*, 1–7. [CrossRef]
13. Hládek, D.; Staš, J.; Pleva, M. Survey of Automatic Spelling Correction. *Electronics* **2020**, *9*, 1670. [CrossRef]
14. Šantić, N.; Šnajder, J.; Dalbelo Bašić, B. Automatic Diacritics Restoration in Croatian Texts. In *The Future of Information Sciences, Digital Resources and Knowledge Sharing*; Faculty of Humanities and Social Sciences, University of Zagreb: Zagreb, Croatia, 2009; pp. 309–318.
15. Ispravi.me Croatian Academic Spellchecker. Available online: https://ispravi.me/ (accessed on 12 April 2023).
16. Mitton, R. Spellcheckers. In *The Routledge Handbook of the English Writing System*; Cook, V., Ryan, D., Eds.; Routledge: Abingdon, UK, 2016; pp. 517–530.
17. Mitton, R. Corpora of Misspellings for Download. Available online: https://www.dcs.bbk.ac.uk/~roger/corpora.html (accessed on 7 April 2023).
18. Toleu, A.; Tolegen, G.; Mussabayev, R.; Krassovitskiy, A.; Ualiyeva, I. Data-Driven Approach for Spellchecking and Autocorrection. *Symmetry* **2022**, *14*, 2261. [CrossRef]
19. Leach, P.; Mealling, M.; Salz, R. RFC 4122: A Universally Unique IDentifier (UUID) URN Namespace. Available online: https://www.rfc-editor.org/info/rfc4122 (accessed on 13 April 2023).
20. Dembitz, Š.; Gledec, G.; Blašković, B. Architecture of Hascheck—An Intelligent Spellchecker for Croatian Language. In *Knowledge-Based and Intelligent Information and Engineering Systems: 14th International Conference, KES 2010, Cardiff, UK, 8–10 September 2010, Proceedings, Part II 14*; Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6277, ISBN 3642153895.
21. Dembitz, Š.; Gledec, G.; Sokele, M. An Economic Approach to Big Data in a Minority Language. *Procedia Comput. Sci.* **2014**, *35*, 427–436. [CrossRef]
22. Banasiak, D.; Mierzwa, J.; Sterna, A. Extended N-Gram Model for Analysis of Polish Texts. In Man-Machine Interactions 5, In Proceedings of the 5th International Conference on Man-Machine Interactions, ICMMI 2017, Kraków, Poland, 3–6 October 2017; Gruca, A., Czachórski, T., Harezlak, K., Kozielski, S., Piotrowska, A., Eds.; Springer: Cham, Switzerland, 2018; pp. 355–364.
23. Ziolko, B.; Skurzok, D.; Michalska, M. Polish N-Grams and Their Correction Process. In Proceedings of the 2010 4th International Conference on Multimedia and Ubiquitous Engineering, Cebu, Philippines, 11–13 August 2010; pp. 1–5.
24. Procházka, V.; Pollák, P. Analysis of Czech Web 1T 5-Gram Corpus and Its Comparison with Czech National Corpus Data. In Text, Speech and Dialogue, In Proceedings of the 13th International Conference, TSD 2010, Brno, Czech Republic, 6–10 September 2010; Sojka, P., Horák, A., Kopeček, I., Pala, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 181–188.

25. Ramasamy, L.; Rosen, A.; Stranák, P. Improvements to Korektor: A Case Study with Native and Non-Native Czech. In Proceedings of the ITAT 2015: Information Technologies—Applications and Theory, Slovensky Raj, Slovakia, 17–21 September 2015.
26. Sorokin, A. Spelling Correction for Morphologically Rich Language: A Case Study of Russian. In Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, Spain, 4 April 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 45–53.
27. Srdić, I.; Gledec, G. Confusion Matrices for Croatian Language. Available online: https://ispravi.me/confusion/ (accessed on 7 April 2023). (In Croatian).
28. Šimunec, M.; Šoić, R.; Vuković, M. N-Gram Based Croatian Language Network. In Proceedings of the 2021 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, 23–25 September 2021; pp. 1–5.
29. Plahuta, M.; Purver, M.; Mathioudakis, M. Gender, Language, and Society-Word Embeddings as a Reflection of Social Ine-Qualities in Linguistic Corpora. Available online: https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/65144/Purver%20Gender,%20Language%20and%20Society%202019%20Published.pdf?sequence=2 (accessed on 12 April 2023).
30. Ulčar, M.; Supej, A.; Robnik-Šikonja, M.; Pollak, S. Slovene and Croatian Word Embeddings in Terms of Gender Occupational Analogies. *Slov. 2.0 Empir. Appl. Interdiscip. Res.* **2021**, *9*, 26–59. [CrossRef]

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.