EasyChair Preprint

№ 11662

Common Typing Mistakes in Kurdish Using "Kurdish Central" Keyboard

Tofiq A Ahmed, Kardo O Aziz, Dilman A Salih,
Ramyar A Teimoor, Abduljabar M Maroof and Harem Kareem

January 2, 2024

Common Typing Mistakes in Kurdish using "Kurdish central" Keyboard

TOFIQ, A, TOFIQ

Computer science, University of Sulaimani, tofiq.ahmed@univsul.edu.iq

KARDO, O, AZIZ

Applied Computer, Charmo University, kardo.othman@charmouniversity.org

DILMAN, A, SALIH

Computer science, University of Halabja, Dilman.salih@uoh.edu.iq

RAMYAR A. TEIMOOR

Computer science, University of Sulaimani, ramyar.teimoor@univsul.edu.iq

ABDULJABAR, M, MAROOF

Professor, Kurdish Department, University of Sulaimani, abduljabar.maroof@univsul.edu.iq

HAREM, KAREEM

Phd, Pasewan Organization, harem@pasewan.com

Automatic detection and correction of spelling errors rely heavily on a statistical study of the trends of spelling errors in a language. In this research project, the spelling error trends for each expert and non-expert in the Kurdish language are determined and analyzed. The statistical study of the error trends is based on data received in real-time from various sources. Traditional (insertion, deletion, transposition, substitution, word separation errors) and language-specific error patterns are found and studied, including position analysis, word length effects, phonetic errors, initial position error analysis, keyboard effects, etc. This is only to check if the Kurdish text is rendered correctly.

CCS CONCEPTS • **Applied computing~Document management and text processing~Document**

Additional Keywords and Phrases: misspelling, Kurdish-Sorani, error pattern, error, typing mistakes

1. INTRODUCTION

Language is the structured medium through which people from the same and different societies and nations communicate. The Kurdish language is spoken by more than 30 million people [1] and used by many computer users, but it lacks linguistic resources and is one of the most underserved languages in the world. Raw texts [2] are the only linguistic resource available on the Internet. However, because they are written in separate scripts, they are insufficient to create a comprehensive and accurate Kurdish corpus, which means that users of the language are denied access to some of the most basic and necessary word processing tools, such as spell check [3]. Other factors affecting the development of the language include a lack of funding, a lack of research, and outdated curricula [4].

Fortunately, Kurdish language research has recently attracted the interest of academics, especially those who are native Kurdish speakers. Google machine translation for the Kurdish dialects of Kurmanji and Sorani, the two main dialects of Kurdish, is one of the most recent developments. The International Conference on Kurdish Linguistics, which began as an informal workshop in Bamberg in2013, has subsequently expanded into a regular international conference series. The Kurdish Dialect [5] database, run by the University of Manchester, is the first large-scale dialect survey of Kurdish accessible via the Internet.

This research examines spelling errors in the Sorani dialect, one of the most widely spoken Kurdish languages. There are no word processing tools, such as a spell checker, for the Kurdish language. Advanced knowledge of spelling errors in any language is extremely useful in developing an accurate spell-checking system for that language. The most accurate error correction is statistical, but for low-resource languages like Kurdish, for which statistical data is not available, the rule-based approach as used in [3] is an alternative. The results of this work can be extremely valuable for developing an accurate spell-checking system, as they facilitate and simplify the selection of possibilities for misspelled words.

Two types of errors are machine-generated errors, such as those produced by OCR or sound-to-text systems, and human-generated errors, such as spelling errors when typing words on a computer [6]. The focus of this study is on human-generated errors.

Spelling errors occur when meaningless words or words not in the dictionary are written. Contextual errors occur when a user writes words that are in the dictionary but have a different meaning than the user expects. Our

contribution to this research is to collect and study data that will be used in the future to improve the training and validation accuracy of artificial intelligence models. Five experienced and five inexperienced individuals were involved in the data collection. Included in this data are unintentional spelling and contextual errors.

The remaining sections of the paper are divided into different sections: The presentation of Sorani texts on computers is discussed in section II, the literature review is discussed in III, the proposed methodology is described in IV, and the results and discussion are included in section V. VI concludes the paper.

1.1. Representing Sorani Text On Computers

Kurdish is a less well-endowed language with a variety of varieties and signs, lacking basic language processing skills. It belongs to the Indo-Iranian family of Indo-European languages, particularly to the Northwest Iranian group of Iranian languages [2], [5], [7], and [8].

It is spoken by the Kurds, the Kurdish people in Kurdistan, who live mainly on the borders of Turkey, Syria, Iran, and Iraq [9]. Small populations also live in the former Soviet Union, particularly Armenia and Azerbaijan, as well as significant exile groups throughout Europe, Central Asia, Caucasia, the Middle East, North America, and Oceania [2], [10, 11]. Persian [7] is the language most closely related to Kurdish.

The essential aspect of Kurdish language is its dialect diversity. It is one of the dialect-rich languages and is described by the term dialect continuum [12]. Kurdish dialects are not mutually intelligible, which means that without established bilingualism [13], people with different dialects may have difficulty understanding each other.

Kurmanji and Sorani are the most widely spoken dialects, both in terms of the number of speakers and the degree of standardization. These two dialects are spoken by more than 75% of Kurds and there are significant differences between them for geopolitical reasons, including morphological and written differences [12].

2. LITERATURE REVIEW

In most languages, the detection of spelling errors has been extensively studied. Several studies have been conducted to investigate the different types and tendencies of spelling errors. Several studies have been introduced on next-word prediction [14] and spelling verification [15, 16].

The public in Kurdistan has access to a spelling checker for the Kurdish language [15]. They recommend an algorithm that looks for typos and offers one or more word-correction alternatives. They discovered that the algorithm's performance and accuracy had improved as a result of the experimental findings[15]. Additionally, they haven't looked at the frequent mistakes and don't comprehend the mistakes that would happen more commonly. As a result, their findings are slightly less impressive than those of studies that analyze the most frequent linguistic errors and understand error patterns, such as those proposed in [16]. Another research [16] has introduced the Urdu spell checker which detects incorrect spellings using a lexicon search technique. It generates a list of candidate words with correct spellings to remove errors using the edit distance approach. Finally, it uses a hybrid model to rank the recommended words.The Kurdish spell checker [15] and an Urdu spell checker [16] are studied because these two languages' characters and writing styles characters and writing styles of the two languages are quite similar. Additionally, we notice discrepancies in their output, and [23] believes that the method that focuses on the most typical errors and finds error patterns, would be able to identify and suggest better word alternatives for each error in a text.

The most known of all the studies of spell checkers is Damerau's [17], because it is written in 1960s and It was written at a time when computer typing was still relatively new. His initiatives were some of the first ones to correct computer typing. Without considering English error patterns, Damerau [17] has tried to create a method for computer-based spelling error detection. According to Damerau [17], one of the following four types encompasses 80% of the typographical errors.

(1) a single letter insertion.
(2) a single letter deletion.
(3) a single letter substitution.
(4) a single letter transposition.

Additionally, Kukich [6] has researched a variety of techniques for automatically identifying and correcting English spelling mistakes. Additionally, Kukich[6] has investigated mistake types and patterns in order to better comprehend them and make word substitution recommendations. Also, Kukich[6] divided the errors into three groups, starting with (1) typographical errors (2) Mind-related mistakes (3) Phonetic mistakes. Another study, Pollock [18], supports the idea that understanding and analyzing error patterns is crucial to improving the quality

of error detection and correction. In order to uncover statistical features, such as the most prevalent kind of spelling errors, they have gathered and categorized misspelled words in academic and scientific literature.

Meanwhile, Brosh [20] examined and categorize the most typical Arabic spelling mistakes made by 63 students. They emphasize the value of learning frequent Arabic mistakes. According to Brosh [20], it is crucial to examine the most frequent spelling mistakes made in each language in order to develop ways for avoiding them. After using those tactics, they ultimately demonstrate how much the students have improved. Altamimi [21] also conducted a comprehensive review study on spelling errors in Arabic and non-Arabic contexts. They shows the difference error types from native and non-native peoples [21]. They have discovered errors involving both interlingual and intralingual, with interlingual errors primarily being caused by the intervention of the primary or mother language [21],

Dastgheib [22] investigated approaches to the detection and correction of spelling errors, focusing on Persian language. They classified the mistakes into non-word and real-word errors. Then they introduced "Perspell," a spelling program that uses a hybrid scoring system and a language model that is optimized for lexicon to fix both types of errors. Additionally, they emphasize the value of looking into error patterns in each language, which will be helpful for identifying and fixing spelling mistakes. As an alternative to the statistical method for error detection, Naseem [23] uses rule-based methods to identify typical errors in Urdu. However, The most accurate error detection techniques are statistical, but for low resource languages, these techniques are out of the question because data is not available [23]. As an alternative, rule-based methods that take advantage of trends in spelling mistakes are helpful. For the creation of such techniques, the study of error patterns in a language is a necessary requirement [23]. Because the qualities of the Urdu language and those of the Kurdish language are so close, our methodology in this work is comparable to that of Naseem [23]. Both languages have limited resources, and the majority of the fundamental tools and data for language processing are not available.

3. METHODOLOGY

Error-free data collection is a difficult task. To collect data, we selected an error-free base text and asked different authors to retype it. 50 news pieces, blogs, and other texts collected from Kurdsat (https://www.kurdsatnews.com) and Kurdsat-news (https://www.kurdsat.tv). The data was provided by the official persons from the websites, who certified that it had been updated and reviewed to make sure there were no typos. They are typed using a formal language and the Kurdish central keyboard layout. The length of the content varied from 130 to 180 words. And the number of sentences varied between 6-10 sentences in each news. In total, we have 353 sentences and 8130 tokens.

In addition, five experienced and five inexperienced typists were used to retype each message on a computer. Experienced typists are someone who has prior experience as a typist (in Kurdish-Sorani) or data entry for more than 5 years for a range of companies or organizations. They have a typing speed of 50-80 words per minute. Thus, they were faster and had to know keyboard layouts better than novice typists. Inexperienced typists were people with no prior experience as typists and their daily job is not writing on computers. They were not very familiar with the Kurdish keyboard layout. The first rule they established was that they should not go back over an article to correct errors after it was completed. The second rule was that they should use the "Kurdish Central" keyboard layout, a layout that is only for Kurds. We were able to collect 500 retyped messages, most of which were definitely misspelled. In total, we had 79,700 retyped words. The framework diagram is presented in Figure 6.

Error detection and correction were done manually on the computer. Because the Kurdish language is new to computer users and does not have many users like common languages such as English, Arabic, and Spanish, there isn't a standard tool for typo detection and correction. Thus, we developed a web application system in Laravel framework to read the re-typed articles and allow users to select words as typing-error manually by the authors and allow them to enter each typo correction in a field next to each error. The annotators were the authors. We have done the error detection and correction manually on the system. We had permission on the system to detect, correct errors, and review other annotators' work. We managed to review each-others work on the system to make sure no detection or correction was passed by mistake.

Errors were divided into words with only one error and words with multiple errors [24], where words with only one error contained only one spelling error (e.g., company typed as *copany*). Words with many spelling errors are called multi-error words. (For example, a company may be typed as *copnay*).

Typos are classified into four categories [17]:

- Insertion: An extra character is added to the word.

- Deletion: a character is removed from a word
- Substitution: one character is replaced by another.
- Transposition: when two adjacent characters in a word are interchanged.

We go further into error types by dividing each type by the number of times the same error occurs in the word. As a result, each type has a single occurrence and multiple occurrences with the same word. For example, we marked the word "company" as multiple occurrences if it was written as "*ccompaany*", because it has "Insertion" error multiple times, and so on for other types. We also compared the number of common spelling errors between experienced and inexperienced typists. Lastly, we compare the typos based on the word length to see which word length has more errors with respect to the total number of words with the same length.

4. FINDINGS AND DISCUSSION

A study was conducted to collect typed Kurdish texts. The original data includes 50 articles. All data were collected from different Kurdish websites. Then, we selected 5 experienced and five inexperienced authors to retype the articles without checking them for spelling errors. Then, the retyped messages were checked for spelling errors. The results of the harvesting are shown in Figure 1.

There are a total of 6782 misspelled words. Some of the words had many errors, but the majority of the terms had only one. We found that "Deletion" and "Substitution" are the two most common forms of mistakes that occurred in our data. The total number of errors per types are shown in Table 6. In English, we came very close to the results of Kukich [6] in 1993. According to Kukich [6], single errors account for 80 percent of misspelled words in English, while single errors account for 89 percent of misspelled terms in Kurdish. There are 6094 words with a single spelling error and 688 words with multiple typos.

Another result is the number of frequent errors about the length of the corresponding form of the word, as shown in Figure 1. We found that words with a length of 5 to 10 characters contain more frequent spelling errors than words with other lengths. Figure 1 is symmetrical concerning the y-axis. It simply shows that the longer (5-10 chars) the word is, the fewer errors it contains, and the shorter (5-10 chars) a word is, the fewer spelling errors it produces.

In the next step, we compare the number of errors in each word length with respect to the total number of words in the same length. The result is shown in Figure 5. The results show that the percentage of occurring typos goes higher as the length of the word increases. It shows that however, we have fewer errors (as shown in Figure 1) in the words with a length of 10 or greater but the percentages go higher see Figure 5. It can be due to the fact that long words occur less in the Kurdish language in general. And It is expected that the longer a word is, the more spelling errors will occur.

Obviously, the percentage of errors is quite large. We found that the first character of the word is responsible for 13% of all errors. A similar discovery was made by Punjabi researchers [24]. We found that a beginner is 15.8 percent more likely to make a mistake than an experienced professional. The numbers and details can be found in Table 3.

From now on, we will focus on the four types of errors (addition, deletion, substitution, and transposition). We looked at how often each character was misspelled for each type, down to the character level.

1.1. Addition

Addition errors are those words that have an extra character(s). We reviewed the retyped articles to see which characters were inserted incorrectly. Table 4 and Figure 2 illustrate the results, as well as the number of common mistakes for each character respectively. The data reveals that the most common errors occur at (space,'ی','ا','ه','و').

It was found that there were 515 times spaces between words. Most of the time, a user inserted an additional space within a word unintentionally, causing the word to split into two words. The overall error will rise if we examine both sorts of faults, such as blank spaces and misspelled words. There are 20 additional spaces in the retyped news. The two characters (ی/e/,و/ʊ/) are the second most common errors. Characters, such as (ی/e/,و/ʊ/), exist in both single (و،ی) and double (یی/ee/,وو/oo/) forms in Kurdish. For example, "کیمیایی" has (یی/ee/) as double form and "سنوور" has (وو/oo/) as double. But sometimes، in other words, that is not necessary, typed as double in other words which causes an "Addition" error.

1.2.Deletion

When a user forgets or omits a character within a word, this is known as a deletion typing error. Figure 3 displays the result of deleting errors. The horizontal axis shows deleted characters, whereas the vertical axis reflects the frequency of each character in the article. It can be observed from figure 3 that "و/u/, Space, ی/y/, د/d/" are the most commonly omitted characters. When we combine the results with errors in words with multiple errors, the statistics go much higher because we have 22 space errors, 14 for (و), and 5 errors for the rest of the characters.

Characters with double (یی/ee/, وو/oo/) are more likely to cause deletion errors, as mentioned in the preceding section. Many words have double forms, but users frequently forget to enter them as such. The two characters (و،ی) are treated as separate forms (وو،یی) because they change meanings inside a word if the double is replaced with a single and vice versa. For example, the word کور means "boy" has a single letter و/u/, whereas the term کوور means "arched" and has a double letter وو/oo/.

1.3.Substitution

A substitution error occurs when one character is automatically replaced with another. The consequences are significant because it's like having both an addition and a deletion error. Many of these mistakes are single-letter substitutions. The result is shown as a confusion matrix in Figure 4 when the horizontal characters indicate the correct form and the vertical characters represent the incorrect ones. Some substitutions were not included in the graph since they occurred less than 10 times.

The most common substitution is when ڕ/r'/ is substituted with ر/r/, which occurs 890 times, while we have 88 instances where ر is replaced with ڕ. We discovered that the majority of the instances occurred in the first position of the word. Furthermore, no word in Kurdish Sorani begins with the letter ر/r/. The initial character of most words that begin with the letter ڕ/r'/ is substituted with the letter ر/r/[25]. For example, in most places, the word "ڕاهێنان" which means "training" is used. It starts with ڕ/r'/ but in most cases, it's written as ر/r/ mistakenly.

The second most common occurrence is when two characters of(ل/l/, ڵ/l'/) are replaced unintentionally as ڵ is replaced 419-time switch ل, and there are 24 more occurrences of ڵ inside multiple-error words. Another significant quantity in figure 4 is the 257 occurrences where ی/i/ is replaced with ێ/e/. Both of these characters have two visual forms, but the phoneme is the same. When they are false at the end of a word ی،ێ, and then when they are false in the middle of a word ـیـ،ـێـ. Even when there is a substitution between these two characters, the user can usually read the word correctly. For example, if the word سلێمانی/slemæni/ (which is correct) is written as سلیمانی/slimæni/ (which is not correct), still users will read it as the correct form. This indicates that the user may be unaware of the error since he or she is familiar with the right form. We have found 257 occurrences of substituting ێ with ی and 24 times ێ is written instead of ی.

These numbers will rise when we include errors from multi-error words since there are 92 more times for (ڕ <- ر) and three times for (ر with ڕ). Regarding the characters (ی،و،ر،ل) each one has a visually similar appearance with an extra "V" on it (ێ،ۆ،ڕ،ڵ) and the pronunciations are close to each other (ی/i/, ێ/e/, ر/r/, ڕ/r'/, ل/l/, ڵ/l'/). The keyboard keys for each of these are shown in table 4. Characters with an additional "v" must use the "SHIFT" key or turn on Capslock. As a result, we find that the frequency of missing words with an extra "v" is significantly larger than the number of incorrectly typed characters with a "v" because the user is either too lazy or forgets to press the shift key.

1.4.Transposition

This sort of error occurs when two consecutive character positions are swapped. We found the occurrence of this sort of spelling error to be 38 times out of 6782 which is comparatively less.

2. CONCLUSION

We collected data on the Kurdish language. 10 people participated in this data collection. These data contain different types of errors such as insertion, deletion, transposition, substitution, and word division. We examined the errors of each type as well as the errors produced by individuals in both categories. Since the data is one of the most important tasks for training and validating natural language processing and spell checking. In the future, we plan to use this data to develop a modern spell checker for the Kurdish language's natural language processing.

2.1.Tables

#	Glyph	Code	Shared with Persian	Shared with Arabic
1	ئ	U+0626	✓	✓
2	ا	U+0627	✓	✓
3	ب	U+0628	✓	✓
4	ت	U+062A	✓	✓
5	ج	U+062C	✓	✓
6	ح	U+062D	✓	✓
7	خ	U+062E	✓	✓
8	د	U+062F	✓	✓
9	ر	U+0631	✓	✓
10	ز	U+0632	✓	✓
11	س	U+0633	✓	✓
12	ش	U+0634	✓	✓
13	ع	U+0639	✓	✓
14	غ	U+063A	✓	✓
15	ف	U+0641	✓	✓
16	ق	U+0642	✓	✓
17	ک	U+06A9	✓	✓
18	ل	U+0644	✓	✓
19	م	U+0645	✓	✓
20	ن	U+0646	✓	✓

21	و	U+0648	✓	✓
22	ھ	U+0647	✓	✓
23	ی	U+06CC	✓	✓
24	پ	U+067E	✓	-
25	چ	U+0686	✓	-
26	ڕ	U+0698	✓	-
27	گ	U+06AF	✓	-
28	ڕ	U+0695	-	-
29	ڵ	U+06B5	-	-
30	ڤ	U+06A4	-	-
31	ۆ	U+06C6	-	-
32	ە	U+06D5	-	-
33	ێ	U+06CE	-	-

Table 1: Kurdish Sorani Alphabets

Table 2: Numbers of mistakes for both inexperienced and experienced people for each type

Inexperienced	Type	Experienced	All
3928	Total	2854	6782
616	Addition	410	1026
1475	Deletion	1511	2986
1819	Substitution	913	2732
18	Transposition	20	38
619	1st Position	325	944

Table 3: Number of mistakes for each character that has been added mistakenly a single time within a word

Characters	Total mistakes	Characters	Total mistakes
ا	21	comma	2
space	515	ۆ	3
ی	135	ھ	3
ە	52	ل	2
س	1	ئ	1

پ	1	م	2
و	64	ن	3
ێ	4	ک	3
د	10	ف	1
گ	2	ڤ	1
ر	16	ب	2

Table 4: Characters that have a similar form with an extra ”v” and their keyboard short key

character	Pronunciation	Keyboard short key
ی	/i/	Y
ێ	/e/	SHIFT+Y
ر	/r/	R
ڕ	/’r/	SHIFT+R
ل	/l/	L
ڵ	/’l/	SHIFT+L

Table 5: number of errors with respect to the length of the word and total number of words with the same error.

Word-length	Total number of words	Number of misspelled	Ratio %
1	3770	6	0.1592
2	7094	27	0.3806
3	5304	178	3.356
4	6376	344	5.3952
5	14057	941	6.6942
6	9978	683	6.8451
7	8813	912	10.34835
8	7232	886	12.2511
9	5958	810	13.5951
10	3811	654	17.16085
11	2076	462	22.2543
12	1216	286	23.5197
13	928	212	22.8448
14	621	138	22.2222

15	271	85	31.3653
16	154	54	35.0649
17	59	27	45.7627
18	68	17	25
19	17	15	88.2352
20	13	9	69.23

Table 6: The total number of errors per each type

Error Type	# of errors
Insertion	1011
Deletion	2982
Substitution	2731
Transposition	58
Total	**6782**

1.2.Figures

2.1.1.Number of misspelled words with regard to their length

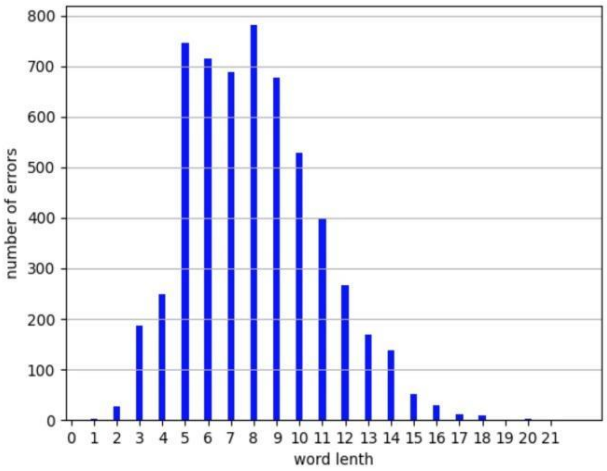


Figure 1: Number of misspelled words with regard to their length

2.1.2.Number of mistakes for each character that has been added mistakenly a single time within a word

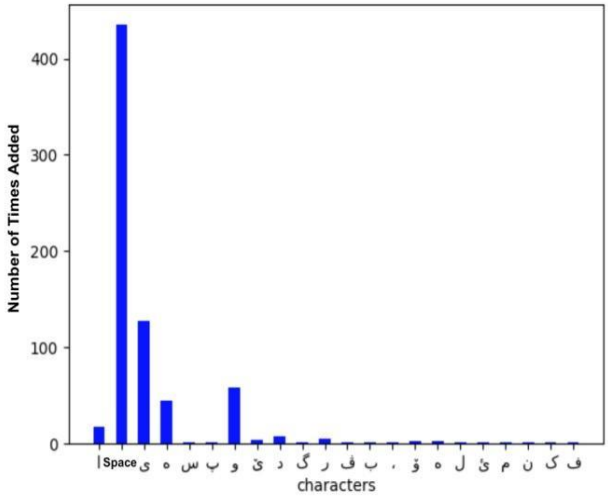


Figure 2: Number of mistakes for each character that has been added mistakenly a single time within a word

2.1.3.Number of deletion times for each character within a word

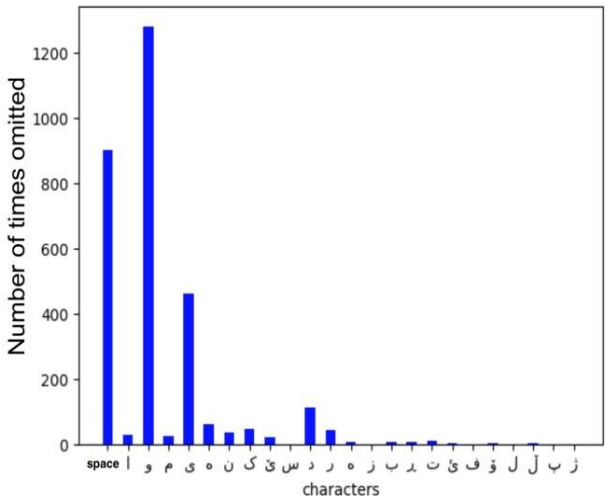


Figure3: Number of deletion times for each character within a word

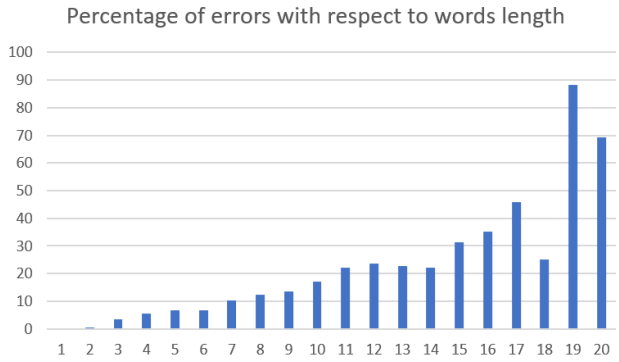


Figure 5: Percentage of error per word length with respect to the total number of words with the same length.

2.1.4.Full-Width Figures.

Figure 4 Substitution confusion matrix, horizontal shows correct characters, vertical shows characters written mistakenly

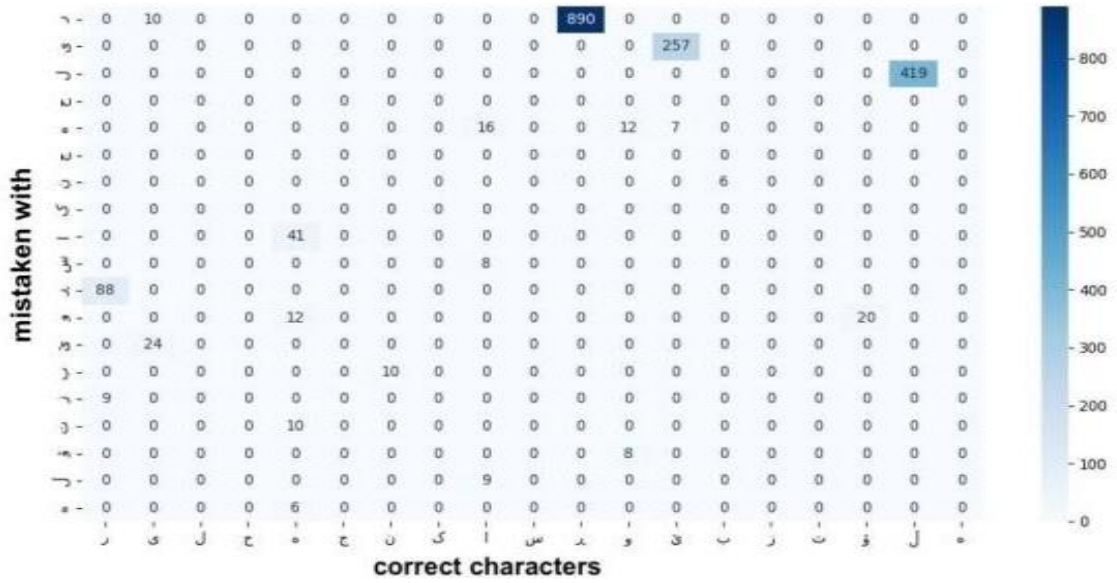
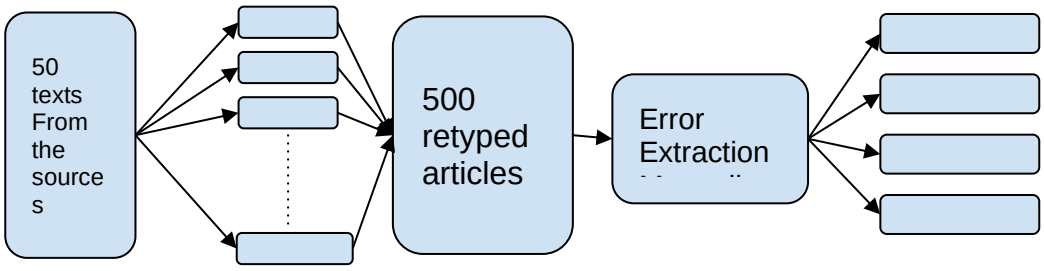


Figure 4: Substitution confusion matrix, horizontal shows correct characters, vertical shows characters written mistakenly.



Step1: Collection original news and articles from the certified source

Step2:Having the writers to re-write the news and articles

Step3: We achieved 500 re-typed articles

Step4:The errors are extracted from each re-typed articles

Step5:Each error classified into one of the four category

Figure 6: The overall framework diagram of the process is shown starting from step1 till step 5

ACKNOWLEDGMENTS

This research was supported by Pasewan Organization. We are grateful to all of those with whom we have had the pleasure to work during this and other related projects. We thank all friends and colleagues for their valuable and constructive suggestions during the planning and development of this research work

1. HISTORY DATES

REFERENCES

1. H. Veisi, M. MohammadAmini, H. Hosseini, "Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus," Digital Scholarship in the Humanities, vol. 35, no. 1, pp. 176–193 (2019 02), [Online]. Available: 10.1093/llc/fqy074, URLhttps://doi.org/10.1093/llc/fqy074
2. . Walther, B. Sagot, "Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish," presented at the Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)(2010).
3. D. A. Salih, Kurdish Sorani Spelling Checker System, mathesis, School of Computer Science, University of Birmingham (2016).
4. B.J.Ali, B.Gardi, B.J.Othman, N.B.Ismael, S. Sorguli,B.Y. Sabir, S.A.Ahmed, P.A.Hamza, H.M.Aziz, G. Anwar, "Educational system: The policy of Educational system in Kurdistan Region in public Kindergarten,"(2021), [Online]. Available: 10.22161/ijels.63.10, URL http://dx.doi.org/10.22161/ijels.63.10 Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (September 1999), 604–632. https://doi.org/10.1145/324133.324140
5. S. G¨undo˘gdu, E. ¨Opengin, G. Haig, E. Anonby,Current issues in Kurdish linguistics, vol. 1 (University of Bamberg Press) (2019).
6. K. Kukich, "Techniques for Automatically Correcting Words in Text,"ACM Comput.Surv.,vol.24,no.4,p.377textbf–439(1992 dec),[Online].Available:10.1145/146370.146380,URL https://doi.org/10.1145/146370.146380.
7. K. Sheykh Esmaili, "Challenges in Kurdish Text Processing," (2012 12).
8. H. Hassani, D. Medjedovic,et al., "Automatic Kurdish dialects identification,"Computer Science & Information Technology, vol. 6, no. 2, pp. 61–78 (2016).
9. A.Mahmudi,H.Veisi,"Automated Grapheme- to-Phoneme Conversion for Central Kurdish based on Optimality Theory," Computer Speech Language vol.70,p.101222(202105),[Online].Available: 10.1016/j.csl.2021.101222.
10. A. Hassanpour, S. Mojab, "Kurdish diaspora in encyclopedia of diasporas," (2005).
11. G. Haig, Y. Matras, "Kurdish linguistics: a brief overview," STUF-Language Typology and Universals,vol. 55, no. 1, pp. 3–14 (2002).
12. K. S. Esmaili, S. Salavati, "Sorani Kurdish versus Kurmanji Kurdish: an empirical comparison," presented at the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 300–305 (2013).
13. J. Sheyholislami, "Language and nation-building in Kurdistan-Iraq," presented at the Middle Eastern studies association 43th annual meeting, Boston, MA(2009).
14. H. K. Hamarashid, S. A. M. Saeed, T. A. Rashid,"Next word prediction based on the N-gram model forKurdish Sorani and Kurmanji,"CoRR, vol. abs/2008.01546(2020), URL https://arxiv.org/abs/2008.01546
15. R. S. Hawezi, M. Y. Azeez, A. A. Qadir, "Spell checking algorithm for agglutinative languages "Central Kurdish as an example"," presented at the 2019 International Engineering Conference (IEC), pp. 142–146 (2019),[Online]. Available: 10.1109/IEC47844.2019.8950517
16. R. Aziz, M. W. Anwar, M. H. Jamal, U. I. Bajwa, "A Hybrid Model for Spelling Error Detection and Correction for Urdu Language," Neural Comput.Appl., vol. 33, no. 21, p. 14707textbf–14721 (2021 nov), [Online]. Available: 10.1007/s00521-021-06110-7, URL https://doi.org/10.1007/s00521-021-06110-7
17. F. J. Damerau, "A technique for computer detection and correction of spelling errors," Communications of the ACM, vol. 7, no. 3, p. 171textbf–176 (1964), [Online].Available: 10.1145/363958.363994
18. J. J. Pollock, A. Zamora, "Collection and characterization of spelling errors in scientific and scholarly text,"Journal of the American Society for Information Science,vol. 34, no. 1, p. 51textbf–58 (1983), [Online]. Available:10.1002/asi.4630340108
19. S.Ahmadi, "KLPT–Kurdish Language Processing Toolkit,"presented at the Proceedings of Second Workshop for NLP Open Source Software(NLP-OSS),pp.72–84(2020Nov.),[Online].Available:10.18653/v1/2020.nlposs-1.11,URL https://aclanthology.org/2020.nlposs-1.11
20. H. Brosh, "Arabic spelling: Errors, perceptions,and strategies,"Foreign Language Annals,vol.48,no.4,p.584textbf–603(2015),[Online].Available:10.1111/flan.12158.

21. D. A. Falah Altamimi, R. A. Rashid, Y. M. Mohamed Elhassan, “A review of spelling errors in Arabic and Non-Arabic contexts,”English Language Teaching, vol. 11, no. 10, p. 88 (2018), [Online]. Available:10.5539/elt.v11n10p88

22. M. Dastgheib, S. koleini, S. Fakhrahmad, “Design and implementation of Persian spelling detection and correction system based on Semantic,”Signal and Data Processing, vol. 16, no. 3, p. 128textbf–117 (2019), [Online].Available: 10.29252/jsdp.16.3.128

23. Naseem, S. Hussain, “Spelling Error Trends in Urdu,” (2007)

24. M. Bhagat, presented at the Single/Multi-Error Misspellings in Punjabi Typed Text (2016)

25. U. (2021, March 8). وانەی حەوتەمی ڕێنووسی کوردیی ئارامی: نەبزوێنەکان. Retrieved January 20, 2022, from https://kurdpress.com/ku/news/1253/