Experience: Insights into the Benchmarking Data of Hunspell and Aspell Spell Checkers

LEENA AL-HUSSAINI, The University of Edinburgh

13

Hunspell is a morphological spell checker and automatic corrector for Macintosh 10.6 and later versions. Aspell is a general spell checker and automatic corrector for the GNU operating system. In this experience article, we present a benchmarking study of the performance of Hunspell and Aspell. Ginger is a general grammatical spell checker that is used as a baseline to compare the performance of Hunspell and Aspell. A benchmark dataset was carefully selected to be a mixture of different error types at different word length levels. Further, the benchmarking data are from very bad spellers and will challenge any spell checker. The extensive study described in this work will characterize the respective softwares and benchmarking data from multiple perspectives and will consider many error statistics. Overall, Hunspell can correct 415/469 words and Aspell can correct 414/469 words. The baseline Ginger can correct 279/469 words. We recommend this dataset as the preferred benchmark dataset for evaluating newly developed "isolated word" spell checkers.

CCS Concepts: • **Applied computing → Document management and text processing**; *Document management*; Text editing;

Additional Key Words and Phrases: Spell checkers, Hunspell, Aspell, Ginger, benchmarking data, error category, error types, data statistics

ACM Reference Format:
Leena Al-Hussaini. 2017. Experience: Insights into the benchmarking data of Hunspell and Aspell spell checkers. J. Data and Information Quality 8, 3–4, Article 13 (June 2017), 10 pages.
DOI: http://dx.doi.org/10.1145/3092700

1. INTRODUCTION

Spell checkers are software packages that check the spelling accuracy of user-typed words. This article presents a benchmark dataset and completes an extensive evaluation of the following two widely known spell checkers: Hunspell [2015] and Aspell [2015]. The benchmark dataset is recommended as the preferred benchmark dataset for newly developed "isolated word" spell checker software.

Hunspell [2015] is a general spell checker and automatic corrector, as well as a morphological (grammatical) checker software. It is used by many popular companies, such as Apache, Mozilla, Google, Apple, Adobe, GNOME, and the X Window System. Examples of applications that incorporate Hunspell are Mac OS X 10.6+ and Safari (Apple), OpenOffice (Apache), Firefox and Thunderbird (Mozilla), Chrome (Google), InDesign and Illustrator (Adobe), gedit (GNOME), and Yudit (X Window System). It is also used by the XML Copy Editor (XML Editor), LyX (document processor), WinShell (IDE for TeX and LaTeX on Windows), and elsewhere (see Wikipedia [2015]).

Aspell [2015] is also a general spell checker and automatic corrector. It is known to be the best English language spell checker. Primarily used in the GNU operating system, and it is also integrated into software applications like Gajim and Notepad++.

Author's address: L. Al-Hussaini; email: ljaerh@aol.com.

DOI: http://dx.doi.org/10.1145/3092700

Table I. Misspelling Statistics

Error Type	**Error Number**	**A**	**B**
Insertion	Single	17	1
Deletion	Single	23	19
Substitution	Single	32	13
Transposition	Single	12	4
Multierrors	4 Positions	1	0
Multierrors	3 Positions	3	1
Multierrors	2 Positions	28	6
	Total	116/117	44/46

Because of the wide popularity of Hunspell and Aspell, this work chose to gain insight into their benchmarking data.

Ginger [2016] is a spell checker and a grammatical checker widely known for its proficiency. It is available as an app for mobile devices, Web browsers, and Microsoft Windows.

The extensive study described in this article will characterize the respective software and benchmarking data from the following different perspectives: error category (e.g., phonetic error), keyboard layout error, author misconceived error, and fast-typo error. This work considers single and multierror statistics and the impact of various length statistics. This work also considers statistics on single error types such as insertion, deletion, substitution, and transposition.

Overall, Hunspell can correct 415/469 words and Aspell can correct 414/469 words. The baseline Ginger can correct 279/469 words. Given the diverse nature of the benchmarking data of Hunspell and Aspell, these 469 words are recommended as the preferred benchmark dataset for evaluating newly developed isolated word spell checkers.

2. MOTIVATION

Although spell checking software has been developed and used for almost six decades, there is not a single work in the literature that proposes a benchmark dataset to evaluate spell checkers. When spell checkers have been evaluated, they appear to have been evaluated on ad hoc datasets. For example, it is common to find the following comments in the literature:

In Damerau [1964]:

> In Table I, the single error technique corrected 87 of the 117 items, and in Table II, 38 of the 46 distinct entries and 71 of the 80 total errors.

From this quote, the author suggests that the software scored 88.75% on single error mistakes. However, we can look further at Table I, which provides more detailed statistics of the results, and make the following two observations:

First, consider the two datasets labeled A and B in the table. One can observe that the sample size is quite small for single error tests. It is noted that it was probably a reasonable dataset in 1964.

Second, although this dataset is available in the ACM digital library, no works have used this data to test their spell checkers. The dataset has also not been updated since 1964.

In Kukich [1992]:

> PZ84 found that only 6% of 50,000 nonword spelling errors in the machine-readable databases they studied were multi-error misspellings. Conversely, Mitton-1987

> found that 31% of the misspellings in his 170,016-word corpus of handwritten essays contained multiple errors.

Looking at multiple errors in a word, one can see from the preceding quotation that both studies used different benchmarking data to find the percentage of multierrors in a word. One study found that the frequency of multierrors was 6%, and the other study found that multierrors comprised 31%. If the author comes with her own test data and tries to find how many of the misspellings are multierrors, she will probably have a different percentage. And if another researcher did the same, then he would most likely come up with a different percentage as well (except in rare cases where there is a closeness in the percentages). It seems worthwhile to establish a unified set of test data as a bedrock to test our solutions without the need to waste time searching for test data and looking at thousands and thousands of words unless there is a good reason for that, which is not our focus. Indeed, looking at 1,000 words and trying to find their statistics and doing a detailed study is really quite a chore, as one might need to scan the list over and over again many times.

From these two motivating cases (among many other cases found in the literature), it is concluded that there is a need for a benchmark dataset to evaluate newly developed spell checking solutions. To be effective, this unified set of data has to be of manageable size and must possess sufficient samples of single errors, multierrors, and other error categories.

There are several reasons the benchmark dataset described in this work (available at: http://aspell.net/test/cur/) is recommended as a preferred benchmark dataset for evaluating spell checker software. Note that there are 517 isolated words in the dataset, of which 469 words have been chosen for the benchmark.

The reasons for the recommendation are as follows:

(1) They are carefully chosen to be of very bad spellers in which the errors challenge any spell checker software. Examples include cumba → combo, develpond → development, and disssicion → discussion, among others. Note that these 469 isolated words are high in phonetic errors, but these errors are challenging and not common misspellings. They are of especially bad spellers.

(2) Even though this data consists of only 469 words, its quality in challenging spell checkers is high. Quality, in our case, matters more than quantity.

(3) Wikipedia [2016] also has a list of common misspellings. As the name states, *common misspellings*, rather than bad misspellings, are more challenging for spell checkers. To prove that, 4,235 words were extracted and tested to see if they contain phonetic misspellings using the double metaphone algorithm [Philips 2000]. They appeared to have 3,201 phonetic errors, comprising about 75.58% of the whole dataset. Thus, it is inferred that this list is good for basic phonetic error testing and not for general spell checker testing, as it is not challenging from different perspectives. In addition, there are redundant errors propagated throughout different variations of a word. For example, this list includes the related misspellings *accomadate*, *accomadates*, *accomadating*, *accomadation*, and *accomadations*.

(4) Dyslexic people have their own list of common misspellings (word list can be found in the link provided in the appendix). Again, as the name suggests, these misspellings are common and not the particularly difficult misspellings required to challenge newly developed spell checkers. To prove that, 155 isolated words were extracted from this list, and it was found that 129 out of 155 words were phonetic errors using the double metaphone algorithm [Philips 2000]. Additionally, all misspellings were common and not challenging.

3. BACKGROUND

3.1. Hunspell and Aspell

The Hunspell and Aspell benchmarking data are the focus of this article. In this section, an overview is given on the mechanics of each algorithm.

3.1.1. Aspell. Aspell [2015] is a phonetic algorithm that converts each misspelled word to its modified metaphone of the English language. It uses the edit distance methodology in counting the number of insertions, deletions, substitutions, and transpositions between a misspelling and a lexicon. A misspelling with a lexicon replacement appears in the suggestion list. The Aspell score metric is the weighted edit distance. It takes the weighted average of the weighted edit distance of the lexicon to the misspelled word. The word combination (misspelling and lexicon) with the lowest score appears first in the suggestion list.

3.1.2. Hunspell. Hunspell [2015] extends support for language peculiarities—Unicode character encoding, compounding, and complex morphology. Hunspell provides suggestions using n-gram similarity, rule, and dictionary-based pronunciation data. It does morphological analysis, stemming, and generation.

3.2. Isolated Word Error Correction Research

This research [Kukich 1992] concerns detecting isolated word errors that appear in text recognition and text editing applications. Research in this area has expanded from the 1960s to the present. Isolated word error correction research not only detects errors but also gives a list of suggested corrections. An example of an isolated word error is the word *cta* but its correct spelling is *cat*. This word stands alone in isolation, and this research does not investigate the context surrounding the misspelling.

3.3. Error Types Found in the Spell Checking Literature

There are two major classes of errors [Van Berkel and De Smedt 1988]: orthographical error and typographical error. The first class of error, also called *cognitive error*, concerns real-word error, phonetic error, grammar error, and author misconceived error. The second class of error, also called *physical error*, concerns keyboard layout error, typing error, split error, run-on error, transmission error, storage error, and residual error. We explain each error briefly in the following:

(1) Orthographical (Cognitive) Error
  (a) *Real-word error* [Peterson 1986]: Misspelling error that results in a correct word, such as typing *picaresque* for *picturesque*.
  (b) *Phonetic error*: Misspelling error that results in a phonetic word, such as *fonetik* for *phonetic.*
  (c) *Grammar error*: Misspelling error that violates grammar rules, such as typing *rsx* for *RSX*, which violates the capitalization rule.
  (d) *Author misconceived error* [Peterson 1980]: Misspelling error that sounds and looks the same as the correct word, such as typing *anonomy* for *anatomy*.

(2) Typographical (Physical) Error
  (a) *Typing error* [Kukich 1992]: Misspelling error that is a result of fast typing or slow typing. In fast typing, the user usually drops letters due to speed, such as typing *amification* for *amplification*. In slow typing, the user usually repeats letters due to slow typing speed, such as typing *independenent* for *independent*.
  (b) *Keyboard layout error* [Kukich 1992]: Misspelling error that is a result of typing a nearby key of the correct actual key, such as typing *abouy* for *about*.

The following physical errors are due to physical devices:

(c) *Split error* [Kukich 1992]: Device error that splits one word, such as splitting *forgot* to *for got*.
(d) *Run-ons error* [Kukich 1992]: Device error that combines two separate words into one word, such as *of the* to *ofthe*.
(e) *Transmission error* [Levenshtein 1966; Peterson 1980]: Error that usually occurs in channel transmission, such as sending *0* for *1* or vice versa. Additionally, sometimes *nothing* is sent and *0* or *1* is received. An example is sending *cat* and receiving *at*.
(f) *Storage error* [Peterson 1980]: Error that occurs in data storage, such as missing characters of a word or transforming it to another. For example, *cat* → *at*.
(g) *Residual error*: Error that adds to a word in the form of a residue. For example, *cat* → *catqA*. The residue could be alphabetical characters or some other strange characters.

The misspelling errors that this article studies are phonetic, author misconceived, typing, and keyboard layout.

4. HUNSPELL AND ASPELL BENCHMARKING DATA STATISTICS

This section presents the diverse nature of the proposed benchmarking data that makes it applicable to test newly developed spell checkers. Seven features of the data are explored: error category, single and multierror classification, details of single error classification, ranking, details of multierror classification, difference in length between a misspelling and a lexicon statistic, and misspelling length statistics.

This work compares Hunspell, Aspell, and Ginger over 469 misspellings. The 469-word sample is part of a larger sample of 547 [Aspell Tests 2015] misspellings providing a manageable sample size for testing spell checkers. The work is concerned with isolated word errors given in four error categories: phonetic, misconceived, keyboard layout, and fast typing. Real-word and grammar errors are outside the scope of this article.

To reproduce the results of the tables, the author has compiled some information serving the purpose:

(1) *Hunspell*: Hunspell 1.3.3 (released on February 6, 2014) is used. The dictionary version used is the August 29, 2007, release. The misspellings are checked one by one from a command-line prompt by moving into the en_US dictionary and typing the following: en_US> hunspell misspellings.txt <enter>. Later, Hunspell will correct them one by one for 469 words. Note that a newer version of Hunspell was released recently, but it is not yet in stable form and thus is not used.
(2) *Aspell*: Aspell 0.60.6.1 (released in July 2011) is used. The dictionary is the U.S. English version 2015.04.24-0. To check the misspellings, in a command-line prompt, > aspell -c misspellings.txt <enter> is typed. Then the 469 words, one by one, are checked.
(3) *Ginger*: The latest version of Ginger(3.7.95) is used. The dictionary is coupled within the software and is not separately available. The misspellings are manually checked in a text document.

The three most popular spell checkers were used: Hunspell, Aspell, and Ginger. These softwares are freely available online. The inclusion of Microsoft Office 2013 was considered, but it requires purchasing or renting, which is not good for ease of reproducing the results and thus is excluded. In addition, the Google (Did you mean?) technology spell checker was considered; however, due to its large dictionary size, it created lots of real-word errors that are not considered in this work and thus is excluded. Microsoft Office 365, which is available online, was tried as well, but it

Table II. Classification of Data According to Error Categories

Error Category	Error Type	Total	Hunspell Can Correct	Aspell Can Correct	Ginger Can Correct
Orthographical	Phonetic	361	327 (90.58%)	336 (93%)	246 (68.14%)
	Author misconceived	68	50 (73.52%)	42 (61.76%)	16 (23.52%)
Typographical	Keyboard layout	11	11 (100%)	11(100%)	6 (54.54%)
	Fast typing	29	27 (93.10%)	25 (86.20%)	11 (37.93%)
	Total	469	415 (88.48%)	414 (88.27%)	279 (59.48%)

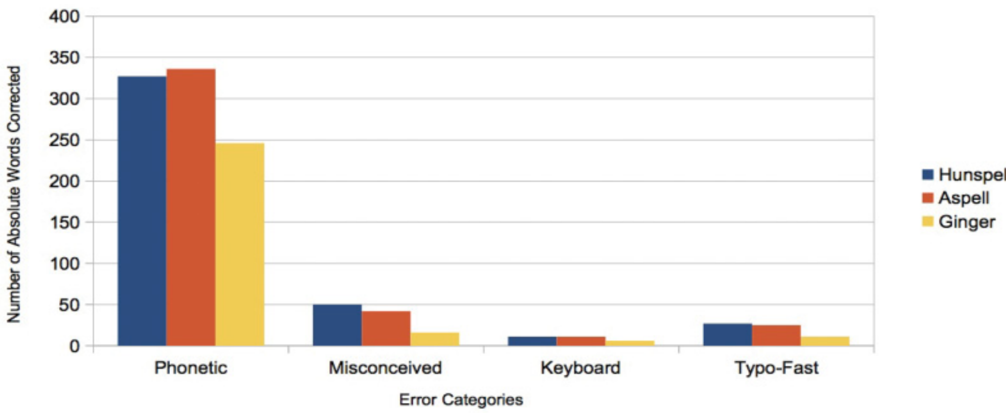

Fig. 1. Single and multierror data statistics.

creates lots of real-word errors, which makes it discouraging to include in the testing of spell checkers. Thus, Hunspell, Aspell, and Ginger are the only ones considered. For the benchmarking data, only words available in all three software dictionaries are considered. Therefore, the remaining 469 words are considered as the benchmarking data to test the three spell checkers.

From Table II and Figure 1, one can observe that Hunspell and Aspell benchmarking data cover errors from both error categories, namely Orthographical and Typographical. In addition, the benchmarking data contains the highest number of phonetic errors, followed by user misconceived errors, fast-typing errors, and keyboard layout errors.

The judging rules used to count phonetic errors in Table II and the other error types are as follows:

(1) An error is phonetic if the double metaphone [Philips 2000] algorithm rules can be applied.
(2) An error is keyboard layout if it contains a wrong character that is near a correct character in the keyboard.
(3) An error is fast typing if there exists one or more deletion(s), one transposition, or one extra letter.
(4) An error is author misconceived if the error does not satisfy any of the preceding rules.

One can observe from Table II and Figure 1 that both Hunspell and Aspell are capable of correcting phonetic errors at a high rate, with Ginger at a lower rate. Overall, Hunspell achieves 88.48%, Aspell achieves 88.27%, and Ginger achieves 59.48%.

Table III shows that the 469-word sample contains 250 single errors and 219 multierrors. For both single errors and multierrors, Hunspell and Aspell scored higher than Ginger. For multierrors, Hunspell corrected 76.25%, Aspell corrected 75.79%, and Ginger corrected 44.29%.

Table III. Classification of Data According to Single Error and Multierror

Error Type	Total	Hunspell Can Correct	Aspell Can Correct	Ginger Can Correct
Single error	250	248 (99.2%)	248 (99.2%)	182 (72.8%)
Multierrors	219	167 (76.25%)	166 (75.79%)	97 (44.29%)
Total	469	415 (88.48%)	414 (88.27%)	279 (59.48%)

Table IV. Ranking Records of Spell Checkers

Rank	Hunspell Can Correct	Aspell Can Correct	Ginger Can Correct
1	276	258	279
2	53	38	0
3	39	48	0
4	16	21	0
5	16	24	0
6	5	10	0
7	5	2	0
8	2	5	0
9	1	1	0
10	1	7	0
11	1	0	0
Total	415	414	279

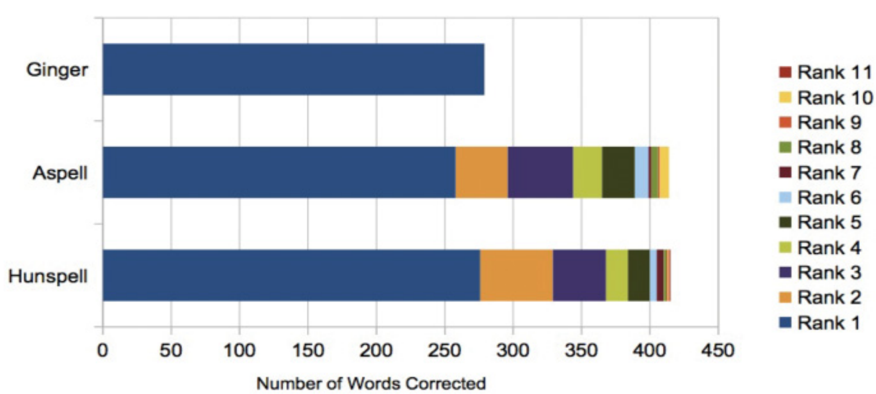

Fig. 2. Ranking records of Hunspell, Aspell, and Ginger.

Table IV and Figure 2 show the ranking records of the three spell checkers. Ginger and Hunspell have higher first ranking records than Aspell.

Table V gives the detailed statistics for a single error. There are four types of errors as mentioned in Damerau [1964] and Peterson [1980]: insertion (addition of a single character in a word, such as *cat* → *catq*), deletion (omission of a single character from a word, such as *cat* → *at*), substitution (exchange of one character with another, such as *cat* → *cst*), and transposition (interchange of two adjacent characters in a word, such as *cat* → *cta*).

From Table V, the 469 sample words contain 250 misspellings of single error type with a mixture of the preceding error types (insertion, deletion, substitution, and transposition.) From the statistics, one can observe that the sample contains the highest number of single substitution errors, followed by deletion errors, insertion errors, and transposition errors. Both Hunspell and Aspell can solve single errors at a very high percentage.

Table V. Classification of Data According to Single Error Type

Single Error Type	Total	Hunspell Can Correct	Aspell Can Correct	Ginger Can Correct
Insertion	62	62 (100%)	62 (100%)	52 (83.87%)
Deletion	69	69 (100%)	69 (100%)	52 (75.36%)
Substitution	93	91 (97.84%)	91 (97.84%)	62 (66.66%)
Transposition	26	26 (100%)	26 (100%)	16 (61.53%)
Total	250	248 (99.2%)	248 (99.2%)	182 (72.8%)

Table VI. Multierror Statistics

Multierror Types	Total	Hunspell Can Correct	Aspell Can Correct	Ginger Can Correct
2 Errors in a word	120	97 (80.83%)	108 (90%)	67 (55.83%)
3 Errors in a word	58	46 (79.31%)	40 (68.96%)	25 (43.10%)
4 Errors in a word	28	18 (64.28%)	13 (46.42%)	2 (7.14%)
5 Errors in a word	10	5 (50%)	4 (40%)	3 (30%)
6 Errors in a word	3	1 (33.33%)	1 (33.33%)	0 (0%)
Total	219	167 (76.25%)	166 (75.79%)	97 (44.29%)

Table VII. Difference in Length Between Misspelling and Lexicon Statistics

Difference of Length	Total	Hunspell Can Correct	Aspell Can Correct	Ginger Can Correct
0	194	176 (90.72%)	179 (92.26%)	113 (58.24%)
1	208	193 (92.78%)	196 (94.23%)	144 (69.23%)
2	51	37 (72.54%)	31 (60.78%)	21 (41.17%)
3	13	7 (53.84%)	7 (53.84%)	1 (7.69%)
4	2	2 (100%)	1 (50%)	0 (0%)
5	1	0 (0%)	0 (0%)	0 (0%)
Total	469	415 (88.48%)	414 (88.27%)	279 (59.48%)

Detailed statistics on multierrors are given in Table VI. It is clear from the data that two errors in a word are much more common than all other multierrors. Second are three errors in a word, followed by four errors in a word, five errors in a word, and six errors in a word. Aspell can correct two errors in a word at a high rate of 90%, whereas Hunspell scored 80.83%. For three to four errors in a word, Hunspell scored higher than the other spell checkers. Overall, Hunspell scored 76.25%, Aspell scored 75.79%, and Ginger scored 44.29%.

Some of the other worthwhile statistics about the benchmarking data are the following:

(1) The difference in length between a misspelling and a lexicon is recorded in Table VII. A difference of length one is the most common in the benchmarking data of around 208 words. Next is the difference of length zero in 194 words. A difference of length of two is reasonably common at 51 words, whereas a difference of lengths three, four, and five are quite infrequent by comparison.
(2) Misspelling word length statistics are recorded in Table VIII. A comparison between Hunspell, Aspell, and Ginger is also recorded. One can see that misspelling lengths ranging from 5 characters to 11 characters dominate the length size statistics, for a total of 404 out of 469 words.

5. RECOMMENDATION FOR BENCHMARKERS

The tables presented in this article give different insights into the benchmarking data of Hunspell and Aspell. This data is worthy of being benchmarking data for newly developed spell checkers. Any table can be used for benchmarking one's

Table VIII. Misspellings Word Length Statistics

Misspelling Word Length	Total	Hunspell Can Correct	Aspell Can Correct	Ginger Can Correct
15	1	1 (100%)	1 (100%)	1 (100%)
14	2	2 (100%)	2 (100%)	1 (50%)
13	12	11 (91.66%)	11 (91.66%)	9 (75%)
12	24	23 (95.83%)	23 (95.83%)	14 (58.33%)
11	43	39 (90.69%)	38 (88.37%)	31 (72.09%)
10	63	59 (93.65%)	56 (88.88%)	46 (73.01%)
9	55	53 (96.36%)	46 (83.63%)	39 (70.90%)
8	69	60 (86.95%)	59 (85.50%)	43 (62.31%)
7	74	61 (82.43%)	65 (87.83%)	45 (60.81%)
6	50	38 (76%)	42 (84%)	22 (44%)
5	50	45 (90%)	48 (96%)	19 (38%)
4	18	15 (83.33%)	15 (83.33%)	6 (33.33%)
3	8	8 (100%)	8 (100%)	3 (37.5%)
Total	469	415 (88.48%)	414 (88.27%)	279 (59.48%)

solution depending on one's focus. There is also an opportunity to compare one's spell checker to the world's best spell checkers of leading software like Macintosh, GNU, and Ginger.

6. CONCLUSIONS

This dataset is both diverse and optimally sized for effective use as benchmarking data to test newly developed spell checkers. In addition, several different ways are shown to measure the data that can be used for evaluation, such as misspelling length, error position, and difference of length between a misspelling and a lexicon. A comparison between Hunspell, Aspell, and Ginger was also recorded given the benchmarking data.

ACKNOWLEDGMENTS

The author is grateful to her Lord for his help in writing this article, first and foremost. The author also would like to thank Kevin Atkinson (Aspell author) and Laszlo Nemeth (Hunspell author) for their replies on their algorithm strategies. In addition, the author would like to thank Emily Paraskevas, Charles Price, and Professor John Grant for proofreading the article. The author would like to thank the reviewers for their valuable feedback and comments, which dramatically improved the work.

REFERENCES

Aspell. 2015. A.1 Aspell Suggestion Strategy. Retrieved May 30, 2017, from http://aspell.net/man-html/Aspell-Suggestion-Strategy.html.

Aspell Tests. n.d. Home Page. Retrieved May 30, 2017, from http://aspell.net/test/cur/batch0.tab.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7, 3, 171–176.

Ginger. 2016. Home Page. Retrieved May 30, 2017, from http://www.gingersoftware.com.

Hunspell. 2015. Home Page. Retrieved May 30, 2017, from http://hunspell.github.io.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys* 24, 4, 377–439.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 8, 707.

James Peterson. 1980. Computer programs for detecting and correcting spelling errors. *Communications of the ACM* 23, 12, 676–687.

James Peterson. 1986. A note on undetected typing errors. *Communications of the ACM* 29, 7, 633–637.

Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ Users Journal* 18, 6, 38–43.

Brigitte van Berkel and Koenraad De Smedt. 1988. Triphone analysis: A combined method for the correction of orthographical and typographical errors. *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLC'88)*. 77–83.

Wikipedia. 2015. Home Page. Retrieved May 30, 2017, from http://www.wikipedia.com.

Wikipedia. 2016. Wikipedia: Lists of Common Misspellings/For Machines. Retrieved May 30, 2017, from https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines.

Received October 2015; revised January 2017; accepted January 2017