Evaluating Evaluation Metrics for Spelling Checker Evaluations

Gerhard B. van Huyssteen*, Roald Eiselen & Martin Puttkammer
Centre for Text Technology, North-West University, Potchefstroom, 2531, South Africa
Tel: +27 18 299 1488; Fax: +27 18 299 1562
E-mail: {afngbvh; ntlere; ntlmjp}@puk.ac.za
*Corresponding Author

Abstract

In the increasingly competitive proofing tools market, it is becoming ever more important to find evaluation methods and metrics that provide stable, invariable measurements. This article focuses on the evaluation metrics currently used in spelling checker evaluation. We commence with a brief explication of some presuppositions relevant to the evaluation of spelling checkers. We then give a detailed description of metrics currently used in various evaluations of proofing tools, discussing and demonstrating their shortcomings. Subsequently, we illustrate that, by adjusting some of these metrics, by incorporating additional information (such as the percentage of errors in the text, and a measurement of the suggestion adequacy), and by adapting the evaluation methodology, a single metric can be designed that evaluates the overall linguistic performance of spelling checkers effectively. In conclusion, we suggest how these metrics can be further improved in future research, in order to open up the possibility to set a single benchmark that can be used for any spelling checker, irrespective of the language being evaluated, or the nature of the text used for the evaluation.

1. Introduction

Standard metrics for the evaluation of the linguistic performance of spelling checkers, like lexical and error recall, and precision, have been widely used for many years [2, 5, 6, 7, 9]. Yet, from a usage-based point of view, there still seems to be some shortcomings in these evaluation parameters, mainly due to the fact that not all variables inherent to spelling checker evaluation are represented in these metrics. In this paper we would like to address some of these issues and come up with evaluation metrics that are supplementary to, or refinements of the current metrics.

One of the major shortcomings in current metrics used in the evaluation of spelling checkers is their inability to provide stable, invariable measurements (see Section 3.5 below). Ideally, the performance measure of a particular spelling checker must be constant over a number of evaluations, irrespective of the percentage of mistakes in different texts, the level of difficulty of the texts, the length of the texts, etc. As EAGLES [1] states: "… a metric is reliable inasmuch as it constantly provides the same results when applied to the same phenomena."

In order to resolve this shortcoming, we are of opinion that both the *evaluation methodology* and the *evaluation metrics* could be modified to render a more accurate representation of actual spelling checker performance. This article concentrates on the evaluation metrics used in spelling checker evaluation, and not so much on the evaluation methodology. However, future research should also be directed at re-evaluating evaluation methods, taking into account different usage-based aspects of spelling checkers (e.g. the functionality of a spelling checker to automatically correct frequently occurring mistakes, the genre of texts used as evaluation texts, etc.).

We will commence with a brief explication of some presuppositions relevant to the evaluation of spelling checkers. We will then give a detailed description of metrics currently used in various evaluations of proofing tools, discussing and demonstrating their shortcomings. In the next section, we will illustrate that, by adjusting some of these metrics, by incorporating additional information (such as the percentage of errors in the text, and a measurement of the suggestion adequacy), and by adapting the evaluation methodology, a single metric can be designed that evaluates the overall linguistic performance of spelling checkers effectively. In conclusion, we will suggest how these metrics can be further improved in future research, in order to open up the possibility to set a single benchmark that can be used for any spelling checker, irrespective of the language being evaluated, or the nature of the text used for the evaluation.

2. Some Presuppositions

For purposes of this article, we identify some presuppositions that are taken for granted or are left unattested, and which should also be taken into account in future research on the evaluation of spelling checkers.

1. A prototypical spelling checker has the following basic characteristics and/or functionalities:
   a. It should be able to recognise only correct words as correct and flag only incorrect words as incorrect (i.e. not flag correct words as incorrect, or vice versa).
   b. For incorrect words, it should be able to provide suggestions for correction, and when demanded by the end-user, automatically rectify the mistake based on the suggestions.
   c. It should be able to store new words (e.g. novel proper names or idiosyncratic words used by the end-user) in some kind of user dictionary, in order to recognise such words in future spelling checking sessions.
   d. It should be able to ignore (i) some kinds of words (e.g. words with numbers, or words written in capital letters); and (ii) some individual words (e.g. words written incorrectly on purpose).
   e. With regard to frequently occurring mistakes, it should be able to automatically correct such mistakes.

2. Spelling checkers have no contextual knowledge [6], and words are therefore checked without any textual context (i.e. as lists, and not as running text).
3. When evaluating spelling checkers, both the linguistic and the functional performance of the spelling checkers should be evaluated. Linguistic evaluation pertains to points a., b., and (to a lesser extent) e. above, while functional evaluation to points c. and d. above, as well as to the speed, memory usage, size, etc. of the spelling checker. The focus of this article is only on linguistic evaluation.
4. Ideally, evaluation metrics and benchmarks should be universally applicable to any spelling checker, irrespective of the language being evaluated. This will enable users to compare the performance of different spelling checkers for different languages. No further attention is paid to this desideratum in this article.
5. End-user profiling is very important, and should be taken into account when evaluating spelling checkers [6]. In our opinion, end-users' paramount expectation of a spelling checker is to provide them with texts without any spelling errors (i.e. to flag all spelling errors). While this is most important to the end-user, it is also important that the spelling checker should have a high proficiency of the language (i.e. does not flag too many correct words as incorrect). These end-user preferences are not incorporated in the metrics designed in this article, but should be accounted for in further research.
6. As different spelling checkers can be employed differently by end-users (e.g. while typing, or on demand, or in batch-mode), this should be taken into account when comparing different spellers. For purposes of this article we presume that the spelling checker is called on demand, and therefore do not take into consideration the effects (or efficiency) of any automatic correction of erroneous words.

3. Current Metrics

In 1997, in conjunction with an EAGLES task group, the TEMAA project (*A Testbed Study of Evaluation Methodologies: Authoring Aids*) released the details of a standard methodology for the evaluation of spelling checkers, based on ISO 9126 standards [8]. This can be considered the most comprehensive body of work on the evaluation of spelling checkers. Their aim was to set a methodology that would automate the entire evaluation process, as well as standardising metrics for the evaluation of the linguistic performance of spelling checkers.

With regard to the methodology, they mainly used structured word-frequency lists ("base lists") extracted from differing corpora, representative of written language (e.g. for Danish they used a base list consisting of 6,780 words, and which covered 82% of a corpus of general interest magazines).[1] Secondly, in order to generate lists of erroneous words ("error lists"), they devised an error generating module to generate errors in the base lists (e.g. for Danish, this resulted in an error list of 4,562 incorrect words). These two word lists were used as the evaluation texts for their spelling checker evaluations.[2]

Using these lists, they evaluated the following attributes, which are also commonly evaluated in the evaluation of retrieval systems, syntactic parsers [3], and proofing tools:

- Positive lexical coverage (recall)
- Error coverage (precision)
- Suggestion Adequacy [6]

Most of subsequent evaluations of and research done on spelling checkers have been based on these evaluation metrics, and although some researchers call the different metrics by different names (e.g. [10]), the calculations and methodology are mostly the same. Starlander & Popescu-Belis [7] came up with some refinements on these metrics, as well as some new metrics for their evaluation of proofing tools, which can be accurately implemented in the evaluation of spelling checkers. They distinguish between the following metrics:

- Precision Correct
- Recall Correct (equal to positive lexical coverage)
- Precision Incorrect (equal to error coverage)
- Recall Incorrect
- Predictive Accuracy
- Harmonic mean of the precision of the spelling checker
- Harmonic mean of the recall of the spelling checker

[1] This implies that 18% of the words in the magazine corpus can be considered low-frequency words. If we presuppose that words with a higher frequency are more entrenched, and therefore more likely to be spelled correctly, this approach disregards words that are more likely to be spelled incorrectly (i.e. words occurring less frequently). Also, this approach will not effectively evaluate the capability of a spelling checker to analyse low-frequency morphologically complex words (like novel compounds).

[2] This approach can be somewhat problematic, especially in languages with concatenative compounding (like Afrikaans), since some correct compounds which are generated by means of the corruption rules will not necessarily be in the wordlist used for validating corrupted words. For instance, the valid Afrikaans compound *piesangbome* ('banana trees') might be corrupted by the rules by doubling of the *m*, thereby forming *piesangbomme* ('banana bombs'). Although *piesangbomme* could be a valid word within a given context, it is not a likely word, and in all probability will therefore not be in a wordlist of valid words. Spelling checkers that make use of morphological analysis will not mark this as a mistake, and if it is in the list of errors, the checker will then be unjustly penalised for not flagging the word.

By distinguishing between precision on erroneous and precision on correct forms, and by introducing predictive accuracy and the harmonic mean scores, one gets a much more refined and nuanced view of the actual performance of a spelling checker.

For our proposed metrics, we will combine the metrics of TEMAA [8], Starlander & Popescu-Belis [7], and Van Zaanen & Van Huyssteen [9]. We will categorise these into four broad categories, viz. Recall Measures, Precision Measures (aka Accuracy Measures), Suggestion Measures, and Overall Performance Measures. Following Starlander & Popescu-Belis [7], we distinguish between:

- **True positives** (*Tp*): valid words recognised by the spelling checker, resulting in **correct non-flags**.
- **True negatives** (*Tn*): invalid words recognised by the spelling checker, resulting in **correct flags** (also called “Good flags”).
- **False negatives** (*Fn*): valid words not recognised by the spelling checker, resulting in **incorrect flags** (also called “False flags”).
- **False positives** (*Fp*): invalid words not recognised by the spelling checker, resulting in **incorrect non-flags** (also called “Missed flags”).

Subsequently, we will give a brief overview of these metrics, which are used as our point of departure.

3.1 Recall Measures

Lexical Recall (R_c: Recall Correct – [7]) is defined as the number of valid words in the text that are recognised by the spelling checker (i.e. true positives), in relation to the total number of correct words in the text (i.e. the sum of all true positives and false negatives):[3]

$$R_c = \frac{Tp}{Tp + Fn}$$

The second recall measure is *Error Recall* (R_i: Recall Incorrect – [7]), which is the number of invalid words in the text that are flagged by the spelling checker (i.e. true negatives), in relation to the total number of incorrect words in the text (i.e. the sum of all true negatives and false positives):

$$R_i = \frac{Tn}{Tn + Fp}$$

The ideal for any spelling checker would be, of course, to recognise all valid words as valid, and all invalid words as invalid, scoring 100% on both R_c and R_i. The recall scores are mostly an indication of the comprehensiveness of the lexicon of the spelling checker (i.e. how representative it is of the language), as well as how untainted the lexicon is (i.e. whether the spelling checker lexicon contains any erroneous words). In the case of spelling checkers that incorporate morphological analysis, the recall measures also give an indication of the effectiveness of such morphological analysis.

3.2 Precision Measures

Precision pertains to the flagging accuracy of a spelling checker: how accurate is the spelling checker in assigning non-flags (i.e. to recognise only correct words as correct), and how accurate is the spelling checker in producing good flags (i.e. to flag only incorrect words as incorrect). Starlander & Popescu-Belis [7] call these measures Precision on Correct Words (P_c), and Precision on Incorrect Words (P_i), while TEMAA [8] only uses the latter, and calls it precision. We will follow the more refined view of Starlander & Popescu-Belis [7], and will call the scores Lexical Precision (P_c) and Error Precision (P_i) respectively.

Lexical Precision (P_c) is computed by dividing all correct non-flags (i.e. true positives) by the total number of non-flags (i.e. true positives plus false positives):

$$P_c = \frac{Tp}{Tp + Fp}$$

The number of correct flags (i.e. true negatives) in relation to the total number of flags assigned by the spelling checker (i.e. true negatives plus false negatives) gives an indication of the spelling checker’s *Error Precision* (P_i):

$$P_i = \frac{Tn}{Tn + Fn}$$

Once again, the ideal for any spelling checker would be to score 100% on both Lexical and Error Precision, as the end-user expects of a spelling checker to flag *all* incorrect words, and *only* incorrect words. This will result in a spelling checker that is 100% accurate in the task at hand.

[3] All measures are given as percentages.

3.3 Suggestion Measures

Suggestion adequacy (*SA*) refers to a spelling checker's ability to present the end-user with relevant and accurate suggestions for all true negatives (i.e. incorrect words flagged by the spelling checker). Note that the *SA* of a spelling checker should only be based on true negatives, and not on all negatives, since the aim is to determine how well the spelling checker can suggest corrections for *incorrect words*.

Although *SA* is often not measured in spelling checker evaluations (e.g. [7]), some researchers have already attempted to express the *SA* of spelling checkers as a metric. Paggio & Underwood [6] suggest that the suggestion adequacy of a spelling checker should be determined as percentages of the following four sub-categories:

- The correct suggestion is the first suggestion;
- The correct suggestion is a visible suggestion;
- All visible corrections are incorrect; and
- No suggestions are offered.

Van Zaanen & Van Huyssteen [9] devised a scoring system, using the following scores per instance, which are then added together and divided by the true negatives (*Tn*) to get a score for suggestion adequacy:

- Correct suggestion within first three suggestions = 1
- Correct suggestion anywhere else in suggestion box = 0.5
- No Correct suggestion = 0

These two approaches can be combined successfully by using the sub-categories of Paggio & Underwood [6], and the scoring system of Van Zaanen & Van Huyssteen [9]. We suggest the following scoring system:

- Correct suggestion is the first suggestion = 1 (CS_1)
- Correct suggestion is a visible suggestion = 0.5 (CS_2)
- All corrections are incorrect = -0.5 (*IS*)
- No suggestions = 0 (*NS*)

Therefore, for each correct suggestion, the spelling checker scores 1, and for each visible correct suggestion 0.5. If the spelling checker offers only incorrect suggestions (thereby helping the end-user out of the frying-pan into the fire), it is penalised with -0.5.[4] However, if the spelling checker does not offer any suggestions, it scores 0 (i.e. for not doing anything, the spelling checker is neither rewarded nor penalised). The *SA* is then determined by summing all the scores (where *S* is a score for a suggestion), and dividing it by the total number (*N*) of true negatives (*Tn*):

$$SA = \frac{\sum_{k=1}^{n} S_k}{N_{Tn}}$$

Of course, the ideal spelling checker would score 100%, since it is expected from the spelling checker to provide for each invalid word flagged only one (correct) flag.

3.4 Overall Performance Measure

Ultimately, an end-user probably wants to know: How good is the spelling checker that I am using? That is, what is the overall linguistic performance of the specific spelling checker? In the available literature, very little has been done to design a metric to measure the overall linguistic performance (*OLP*) of spelling checkers. Starlander & Popescu-Belis [7], however, measure the Predictive Accuracy (*PA*) of a spelling checker (i.e. the likelihood of any given word, correct or incorrect, being handled accurately by the spelling checker)[5], and is calculated as follow:

$$PA = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$$

This gives a good overall view of the competency of a checker, since it determines how accurate a spelling checker is at performing the errand it was sent on. By means of this metric, one can express what the spelling checker does right (i.e. the sum of true positives and true negatives) in terms of everything the spelling checker does (i.e. all true and false positives and negatives). Like all the other scores, this score can also be represented as a percentage value, where 100% would be the ideal.

The problem with this measure is that the difference between an adequate spelling checker and an inadequate one is relatively small. Compare for instance Table 2, where three different Afrikaans spelling checkers are compared with each other. Although most of the scores (especially P_c and *SA*) clearly and unequivocally point at

[4] We have decided that the spelling checker should only be penalised with -0.5 (and not -1, for instance), since the end-user still has a choice to accept the suggestion or not. When automatic correction of spelling mistakes is evaluated, a penalty score of -1 could be awarded to wrong corrections.

[5] It is not quite clear why Starlander & Popescu-Belis [7] call this Predictive Accuracy. In our view, this metric does not really make any predictions about the performance of a spelling checker, inasmuch as it rather reflects the performance of a spelling checker in a given evaluation.

Speller C as the best spelling checker, it is more difficult to reach the same indisputable conclusion on basis of the P_a scores only. This means that using the P_a measure as benchmark is very difficult, since very little room is left for nuanced fluctuations inherent to the evaluation of proofing tools.

Two other measures used by Starlander & Popescu-Belis [7] are functions that give a harmonic mean between the recall and precision measures, as calculated above. The overall recall and precision measures on correct and incorrect words respectively are measured, not as an average, but rather as a mean average between 0 and 1 (where 1 is ideal). These measures for correct words (fm_c) and incorrect words (fm_i) are calculated as follow:

$$fm_C = \frac{2}{\frac{1}{R_C} + \frac{1}{P_C}}$$

$$fm_i = \frac{2}{\frac{1}{R_i} + \frac{1}{P_i}}$$

These measures entail that checkers that make use of trivial strategies (for example, by allowing all words to be seen as correct to obtain high lexical recall) will be penalised in terms of their error recall. The mean score is an average measure which calculates an average that is lower than the normal average score. The mean tends to be more severe on lower scores, awarding a score which is closer to the lower score than the normal average score would be. As pointed out, these measures are especially valuable to detect trivial strategies.

3.5 Evaluation of Metrics

The metrics discussed above were employed in a comparative evaluation of three different Afrikaans spelling checkers,[6] using three different texts. These texts were randomly selected from the *North-West University Corpus of Electronic Texts: Afrikaans* (*NWUCETA* – a balanced corpus of Afrikaans texts, taken from the world-wide web, e-mail correspondence, electronic study guides, e-bulletin boards, and other electronic texts). Alphabetical wordlists of types (and not tokens)[7] were extracted from each of these texts, and all erroneous words were then identified by an external proofreader.[8] The three spelling checkers were subsequently evaluated manually, using Microsoft Excel 2000®.

In Table 1 a comparison of three different texts, checked by Speller C, is presented, while Table 2 presents the results of the evaluation of three different spellers, using only 1 text (i.e. the same text). It is clear from the data in Table 1 that current evaluation methods and metrics do not comply with the EAGLES specification that a metric should "constantly provide the same results when applied to the same phenomena" [1]. For instance, compare the huge differences among the P_i scores, as well as the fm_i scores. These results alone motivate the need to re-evaluate current best-practices in the evaluation of spelling checkers.

Amongst other comments that can be made about the data in Tables 1 and 2, suffice it to observe three main trends that follow from the data (as well as numerous other evaluations): regarding the size of the test text, the percentage of errors in the test text, and suggestion adequacy (*SA*).

Speller C	**Text 1**	**Text 2**	**Text 3**
Total words	7009	3178	4805
% Errors	3.72	1.26	0.48
R_c	98.61	99.2	98.93
R_i	92.72	92.5	91.3
P_c	99.72	99.9	99.96
P_i	72.02	59.68	29.17
SA	87.00	81.00	77.00
P_a	98.39	99.12	98.90
fm_c	99.16	99.55	99.44
fm_i	81.07	72.55	44.21

Table 1: Comparing performance of Speller A on three different texts

[6] Due to South African legislation regarding product comparisons, the names of the spelling checkers may not be revealed.

[7] It is also possible to use types, since it is presumed that, by using types, a spelling checker is only rewarded or penalised once for its performance on a certain string. However, from a usage-based perspective, we use tokens, since the end-user sees/types/corrects tokens.

[8] Not only erroneous words were marked, but also abbreviations, acronyms, proper names, words from other languages (e.g. English), and words containing numbers.

Text 1	Speller A	Speller B	Speller C
Total words	7340	7340	7340
% Errors	12.89	12.89	12.89
R_c	99.28	98.07	97.14
R_i	95.12	95.42	99.37
P_c	74.30	73.82	94.99
P_i	95.35	87.32	80.23
SA	69.50	63.00	80.50
P_a	95.15	94.37	96.91
fm_c	97.16	96.73	98.25
fm_i	83.52	80.00	86.99

Table 2: Comparing performance of three spellers on the same text

Firstly, it seems as if the size of the test text could have an influence on the results. This is confirmed by numerous other evaluations, where it was found that the size of the test text has an influence on the stability of the results. To verify this hypothesis, we conducted another set of evaluations, where 4 sets of texts, consisting of 5 different texts differing in sizes ranging from 7,000 to 30,000 words, have been evaluated. The aim of these evaluations was to find the size text where the Overall Linguistic Performance (*OLP*) stabilises. Compare the data in Table 3, where the *OLP* was measured using the metric presented in Section 4 below.

Nr of Words	**7,000**	**15,000**	**20,000**	**25,000**	**30,000**
Set 1	92.33	88.83	90.04	90.43	89.73
Set 2	90.49	90.73	89.22	88.65	89.25
Set 3	84.27	87.93	87.11	88.48	88.62
Set 4	87.24	84.24	89.83	90.54	89.44
Standard Deviation	3.56	2.72	1.34	1.11	0.47

Table 3: Effect of text size on Overall Linguistic Performance (*OLP*)

To interpret the data, we only interpreted the standard deviation with the empirical rule, and worked with the sample mean. We accept hat 1% deviation to both sides of the sample mean would be the norm. From the data in Table 3 it is clear that, on 30,000 words, the standard deviation is 0.47%, which means that 2 Standard Deviations will be smaller than 1% to both sides of the sample mean (in terms of the Empirical Rule for Standard Deviations, this would accommodate 99% of all cases). Of course, our sample size is too small to make any definitive claims, but it could be set as a hypothesis for further research. For purposes of this paper, and until further evidence could be obtained, we conclude that one should ideally use a test text of at least 30,000 words, also ensuring that one always has comparable test corpora (at least in terms of the size of the text).

Secondly, from the evaluations it became apparent that Error Precision is subject to the percentage of mistakes in the text: the higher the percentage of errors in a text, the higher the Error Precision. To verify this observation, we evaluated 22 different Afrikaans texts (ranging in text size between 3,000 and 12,000 words). In Figure 1, a scatter chart of these evaluations is presented, indicating the percentage of errors in terms of the Error Precision. The power trend line indicates the relation between the percentage of errors in a text and Error Precision.

This implies that, in order to set an Error Precision benchmark, the percentage of mistakes in the evaluation text needs to be taken into account. For instance, in the TEMAA evaluation of Danish spelling checkers [6], the percentage of errors that were generated was 67.29% (i.e. 4,562 words in their error list, based on 6,780 words in their base list). This is an excessively high percentage of mistakes, unlikely to occur in any usage-based text, and could therefore explain the high Error Precision scores they obtained in their evaluations (see also [5, 7]). In order for the linguistic effectiveness of a checker to be evaluated from an end-user perspective, realistic percentages of errors should be used for the evaluation process. Please note that this does not mean that the precision measure is incorrect, as it accurately measures exactly what it is supposed to. We are merely arguing that, in order to set a benchmark for Error Precision, a more refined metric should be implemented, taking into account the percentage of errors in a given text (see 4. below).

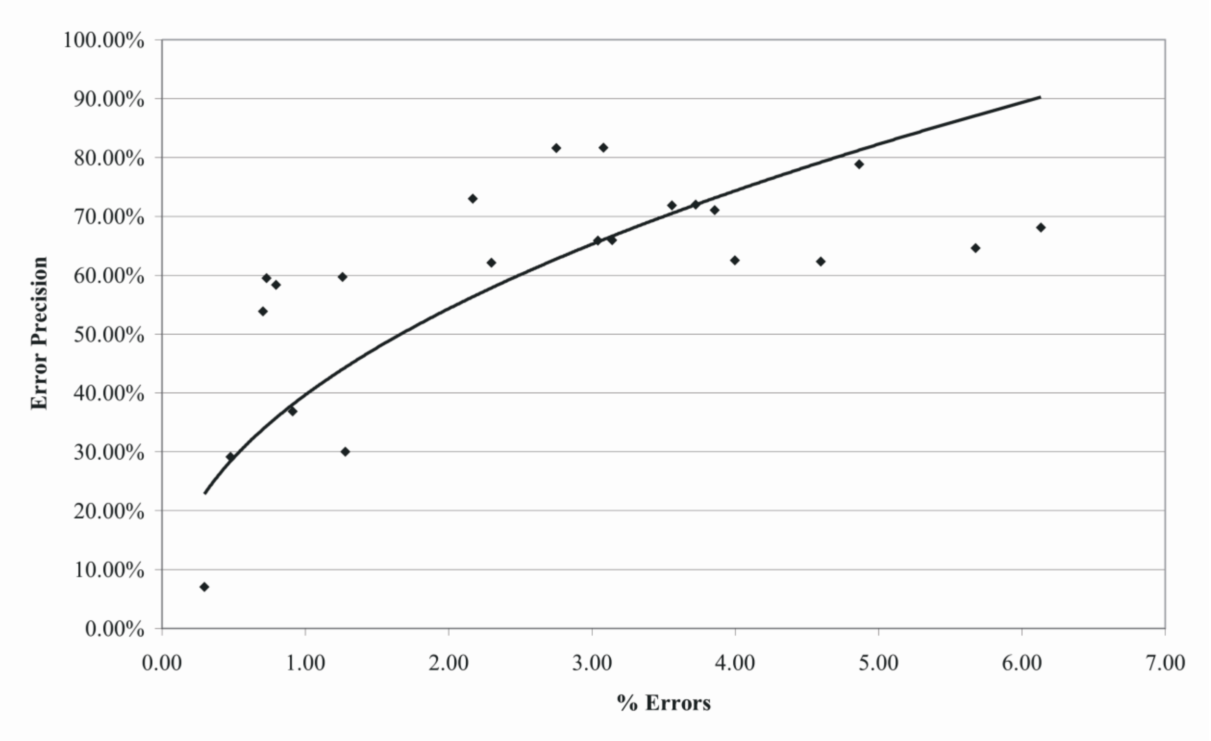

Figure 1: Relation between Error Precision and percentage of errors

Lastly, suggestion adequacy (*SA*) is not taken into account when measuring the overall linguistic performance (P_a) of spelling checkers. However, from a usage-based approach, this attribute should also form part of an overall linguistic performance measure, as it is an important functionality of spelling checkers, specifically for non-native speakers of a language. We are therefore of opinion that suggestion adequacy should be included when determining the overall linguistic performance of spelling checkers.

4. Possible Solutions

Our solution to the problems mentioned above, is to design a new metric, based on existing measures, but incorporating all the different aspects mentioned in Section 3.5 Additionally, we will also suggest an adaptation to the existing methodology. These solutions will prove to provide a more stable and reliable measurement that can be used in evaluating the overall linguistic performance of spelling checkers, and thereby approach the setting of universal benchmarks. This takes into account those aspects that have previously either been ignored, or only described and not actually implemented in the evaluation process.

The first adaptation we suggest is that test texts of at least 30,000 word samples should be used in spelling checker evaluations, since we have indicated above that Error Precision only stabilise when the test text is larger than 30,000 words. Ideally, these texts should be usage-based (i.e. “real” texts created by “real” people, and not frequency lists, or automatically generated/corrupted lists), as well as balanced (i.e. from different genres, modes, sources, etc.).

The second adaptation pertains to the percentage of errors in the texts. We have indicated above that Error Precision (P_i) is a problematic measure, as it deviates considerably, depending on the percentage of errors in the text. P_i should therefore be adapted to provide for the percentage of errors in a text. The ideal way to do this is to set the power line in Figure 1 as a function of P_i. However, due to insufficient data (not only for Afrikaans, but across languages), we rather take the error percentage into account by normalising the amount of errors to a set percentage, thereby at least ensuring that spelling checkers are evaluated equally. The normalisation still takes into account the actual performance of the spelling checker, but levels the playing field for all checkers: if you have a spelling checker that scored poorly on a text with 5% errors, and a spelling checker that has scored reasonably well on 2.6% errors, it will be easier to compare the two spelling checkers when the percentage of errors in the text has been normalised to, say, 10%.

The normalisation percentage is, of course, not that easy to determine. Since the higher this error percentage is, the more stable P_i will become (and therefore of little value as an evaluation metric), we do not want to relate this to a totally unrealistic value. Moreover, we still want to reflect our general usage-based approach to spelling checker evaluation. Based on the percentage errors in our test texts, as well as on Kukich [4], we determined that 6% normalisation on the error percentage is a realistic benchmark to set, since this reflects an average over all unedited texts in our corpus. However, determining the most efficient and accurate normalisation percentage needs to be investigated in future research. This normalised error percentage is then used to calculate the Adjusted Error Precision (P_{ia}), using the following equation:

$$P_{ia} = \frac{Tn\left(\frac{Normalisation\%}{\%ErrorsInText}\right)}{Tn\left(\frac{Normalisation\%}{\%ErrorsInText}\right) + Fn}$$

In order to determine the overall linguistic performance of a spelling checker, we first calculate the overall harmonic mean (fm_o) of all the different precision and recall scores, using the Adjusted Error Precision (P_{ia}):

$$fm_O = \frac{4}{\left(\frac{1}{R_c}\right)+\left(\frac{1}{P_c}\right)+\left(\frac{1}{R_i}\right)+\left(\frac{1}{P_{ia}}\right)}$$

As discussed above, we prefer to use the harmonic mean rather than a normal mean calculation (like P_a), in order to penalise spelling checkers that make use of trivial strategies to obtain high scores [7]. This harmonic mean score also remains more stable when the same checker is evaluated over different texts, unlike P_a.

Subsequently, using this harmonic mean, we calculate the Overall Linguistic Performance (*OLP*) of the spelling checker by incorporating a weighted score of the suggestion adequacy (*SA*) in the metric:

$$OLP = (fm_o \times 0.667) + (SA \times 0.333)$$

The weighting used here (i.e. 67|33) is based on our assumption that recall, precision, and suggestion adequacy are all equally important aspects to determine the overall linguistic proficiency of a spelling checker. However, should further evidence be found that end-users do not place such a high premium on suggestion adequacy, the weights could be adapted accordingly.

Table 4 gives the results of our evaluation of an Afrikaans spelling checker, where the adapted metrics have been employed. Note the relative stability of the overall linguistic performance score.

Speller C	**Text 1**	**Text 2**	**Text 3**
Total words	37328	36843	29267
% Errors in text	3.76	2.58	3.13
% Normalisation	6	6	6
R_c	98.72	98.25	98.53
R_i	96.3	95.59	96.72
P_c	99.85	99.88	99.89
P_i	74.57	59.16	68
P_{ia}	82.39	77.11	80.29
fm_o	93.75	91.7	93.12
SA	81.68	88	81.49
Overall linguistic performance	**89.73**	**88.62**	**89.25**

Table 4: Implementation of new metrics in the evaluation of an Afrikaans spelling checker

5. Conclusion

In this article, we have evaluated metrics currently used in the evaluation of spelling checkers. We have pointed out some shortcomings in these metrics (and evaluation methods), and have indicated ways to improve on these. To this end, we have adapted current metrics and designed some new metrics, which give a more accurate and more stable measurement of the linguistic performance of spelling checkers.

Although these adapted metrics set out above is a step forward in setting a single benchmark metric that is applicable across languages and texts, more research needs to be done on several other aspects. For instance, one of the biggest problems with evaluations stems from the fact that texts differ in their levels of difficulty: results obtained by evaluating a children's books, compared to an evaluation of, say, a physiology handbook will give differing results, even if the texts are of equal size, and error normalisation is applied. The development of a text difficulty score that is relevant to proofing tools (i.e. that can be implemented in the metric for overall linguistic performance), as well as being adaptable for different languages, is therefore needed. These kinds of scores exist for English, but are not necessarily applicable to other languages.

More research also needs to be done on other spelling checking tools such as automatic correction functions, or spelling checkers developed for languages for special purposes. This will be somewhat more difficult to measure, since active participants will have to partake in these evaluations. One also needs to determine how often these tools are used, and what importance is placed on these tools by end-users.

In the increasingly competitive proofing tools market, it is becoming ever more important to find evaluation methods and metrics that are both accurate and unambiguous in their results. To compare the performance of different spelling checkers, even across languages, it is important to set a standard that can be applied in any given context (i.e. on different texts) to determine the efficiency, both linguistic and functional, of proofing tools and more specifically spelling checkers. In the end, this will enable companies and end-users to make better informed decisions when buying proofing tools.

Acknowledgements

We would like to thank Suléne Pilon for helping with the linguistic editing of the test texts, Wannie Carstens for supplying numerous texts for evaluation, and Faans Steyn and Bertus van Rooy for stimulating comments on the statistics.

References

[1] EAGLES. 1995. The EAGLES extension to ISO 9126. Web: [http://www.issco.unige.ch/ewg95/node15.html.] Date: [20 October 2002].

[2] ELSE. 1999. Evaluation in language and speech engineering. Web: [http://www.limsi.fr/TLP/ELSE/FullXreportXver302.htm]. Date: [20 February 2004].

[3] GOODMAN, J. 1996. Parsing Algorithms and Metrics. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL '96)*. pp. 177-183.

[4] KUKICH, K. 1992. Spelling correction for the telecommunications network for the deaf. *Communications of the ACM*. May. 35(5): 80-90.

[5] PAGGIO, P. & MUSIC, B. 1998. Evaluation in the SCARRIE project. *Proceedings of the First International Conference on Language Resources and Evaluation*. pp. 277—282.

[6] PAGGIO, P. & UNDERWOOD, N.L. 1997. Validating the TEMAA LE evaluation methodology: a case study on Danish spelling. *Natural Language Engineering*. 1(1): 1-18.

[7] STARLANDER, M. & POPESCU-BELIS, A. 2002. Corpus-based evaluation of a French spelling and grammar checker. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain. pp. 268-274.

[8] TEMAA. 1997. Evaluation framework for Spelling Checkers: Final Report. Web: [http://www.cst.dk/temaa/D16/d16exp.html]. Date: [20 February 2004].

[9] VAN ZAANEN, MM & VAN HUYSSTEEN, GB. 2003. Improving a Spelling Checker for Afrikaans. In: GAUSTAD, T. (ed.). *Computational Linguistics in the Netherlands 2002: Selected Papers from the Thirteenth CLIN Meeting*. Amsterdam: Rodopi. pp. 143-156.

[10] VERBERNE, S. 2002. Context-sensitive spell checking based on word trigram probabilities. Unpublished MA thesis. Nijmegen: Katholieke Universiteit Nijmegen.