THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

NZQA
NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA
QUALIFY FOR THE FUTURE WORLD
KIA NOHO TAKATŪ KI TŌ ĀMUA AO!

Te Reo Māori Spellcheck Technologies and NCEA examinations

Completed for NZQA March 2021

by Paora Mato, Te Taka Keegan, Dayne Perkins-Gordon
The University of Waikato

1 Introduction

As part of their role, the New Zealand Qualifications Authority (NZQA) are responsible for ensuring the delivery, quality and credibility of New Zealand's secondary school educational qualifications. NZQA are also charged with managing the New Zealand Qualifications Framework (NQF), administering the secondary school assessment system, independent quality assurance of non-university tertiary education providers and qualifications recognition and standard-setting for some specified unit standards.[1]

This research aims to address the lack of Te Reo Māori spellcheck functionality for NCEA digital assessment. Currently, candidates are able to check the spelling of the responses that they have provided in English and, where some words are found to be incorrectly spelt, have correct suggestions provided. The same functionality is not provided for responses provided in te reo Māori which, by comparison, adversely affects candidates responding to digital assessment using Te Reo Māori text. This research will focus on spellchecking for te reo Māori and include observations that may be useful for the implementation of correct word suggestions at some later stage as part of the current software platform employed by NZQA. Additionally, the initial stages of spellcheck functionality for te reo Māori across all relevant NCEA digital assessments should identify and describe other sources of current spellcheck capability for comparison.

Therefore, to fully address the research aims this report will:

1. define the research objectives;
2. provide an explanation of spellchecking technology;
3. identify existing providers;
4. describe how spellchecking can be implemented for te reo Māori using the current spellchecking software (SoNET);
5. provide a comprehensive te reo Māori word list; and,
6. outline any further considerations.

This research is focused on particular aspects of the online delivery and assessment of the National Certificate of Educational Achievement (NCEA), which includes the educational achievement standards normally delivered in New Zealand secondary schools. In an effort to replicate how students already interact with their world, NZQA is supplementing the hand-written NCEA examinations with online digital options. Currently, both paper and digital examinations are able to be answered in either English or te reo Māori, if the exam allows it. Spellchecking and assisted correction are currently available only in the English language.

Current spellchecking functionalities in NCEA digital assessments are not available in te reo Māori, affecting examinations and translations that would use te reo Māori text. Before spellchecking functionalities for te reo Māori can be implemented accurately across all appropriate NCEA digital assessments, research will be completed to identify current spellchecking technologies and to investigate digital sources of te reo Māori suitable for use with this type of functionality.

[1] See: https://www.nzqa.govt.nz/about-us/our-role/

1.1 Rationale (per NZQA):

As an official language of New Zealand, students are able to, and may be required to respond to their NCEA examinations in te reo Māori. The move to digital assessment provides an opportunity to improve the equity of NCEA provision and outcomes for Māori students. The limitations of the current Spellchecking functionality are a hindrance to the NZQA equity goal.

Te Ao Māori is an important principle underpinning the vision to equitably serve Māori students while benefiting the wider assessment system. Digital assessment will ensure an equitable experience for Māori students if a 21st century approach is applied to its development, delivery practices and data analytics. Enabling students to use te reo Māori and mātauranga Māori in their assessments will support the NZQA vision of equitable NCEA outcomes.

Current spellchecking functionalities in NCEA digital assessments are not available in te reo Māori, affecting examinations and translations that would use te reo Māori text. Before spellchecking functionalities can be implemented accurately across all appropriate NCEA digital assessments, research should identify current spellchecking technologies and investigate digital sources of te reo Māori that would be suitable for the deployment of appropriate spellchecking functionality.

1.2 Research aims:

This research aims to identify:

1. existing providers of spellcheck tools or programs that use Te Reo Māori; and,
2. opportunities for developing such functionality in the absence of existing providers.

The list of existing providers of spellchecking tools (see Section 5) is the result of desk-based web searches. The identified tools for te reo Māori have been critiqued and summaries have been included.

A range of programs that use or promote te reo Māori, either in a teaching or in a learning capacity, have been outlined. Many of these have been sourced through our networks either from existing relationships or via word of mouth.

2 Research Overview

Basic Language Resource Kits (BLARK) refer to those types of resources that are able to be used as a platform to deploy advanced end-user language technologies such as:

- monolingual and bilingual corpora
- machine-readable dictionaries
- thesauri
- parts-of-speech taggers
- language parsers
- automatic machine translation
    - text to text
    - text to speech
- speech recognition and synthesis, including text to speech, speech to text
- spellchecking.

Of the world’s 6000-7000 living languages less than 100 enjoy the benefits of these sorts of technologies and the remaining 98+%, often referred to as under-resourced languages, lack many, if not all of these tools.[i] Approximately 5% of the world’s languages are spoken by 95% of the world’s population. Conversely, 95% of the world's languages are spoken by about 5% of the world’s population and are therefore considered to be minority languages.[ii,iii] In many instances the technology build for the lesser-used languages are often constrained by issues of scale because the deployment of such technologies is adversely hampered by a dearth of language data and the lack of capacity for wide-scale uptake and engagement that the larger, more popular languages can provide. In spite of this, particular language groups are investigating an array of processes and models that can be deployed using various methods of statistical analysis and non-annotated training data.[iv,v]

3 Other Considerations

The design and development criteria for spellchecking software for te reo Māori will require suitable corpora to effect accurate checks. An appropriate plan of action, therefore, would include developing a high quality rārangi kupu (word list) either by manually or automatically accruing known sources or by deploying technology that automatically scans the internet (for instance) for suitable words that can then be virtually compiled for word comparison.

Understanding how the current technology employed by NZQA (for English) recognises and matches words will be necessary to ensure any dictionary that is created can be successfully employed by the software currently in use. The culmination of these actions would ideally involve some form of usability testing as proof of concept. Prior to that, vetting by a registered translator would ensure that the initial list would be as accurate and as appropriate for use, especially for proof of concept testing and beyond.

The observed outcomes from the usability testing, aside from confirming suitability, might then result in considerations that could be applied to word suggestion and replacement for te reo Māori at some later point.

4 Spellchecking Technology

Initial spell checking technologies were used with the early word processors and designed to "verify" rather than "correct". Verification software tends to be of limited use since they do not offer assistance in the form of correct word suggestions or replacement for incorrectly spelled words. However, they do signal that a word may be incorrectly spelled.

Current basic spell checking technology has two steps:

1. Error detection
2. Error correction

Error detection involves matching words to a dictionary or list and is generally divided into two categories:

1. Typographical or non-word
2. Cognitive or real word

Typographical errors normally occur by mistake. They do not follow linguistic criteria since they are most often the result of input error (mistyping).[v] Cognitive errors include words that are spelt correctly and are often pronounced similarly to the intended word and/or have similar spellings. These types of words are more difficult to detect but remain contextually incorrect. For instance, *a here ran past me as I was standing hare.* Standard spellcheckers would not catch the errors in this sentence. However, capturing these sorts of errors can be addressed by particular algorithms that use probability techniques, n-grams, and/or neural networks in ways that identify preceding and following words to derive accuracy based on content and context. More sophisticated spell checkers now recognise simple grammatical errors using these methods but, even so, still rarely catch every error contained in a text. However, since word suggestion and replacement is not a part of this project scope, the research will focus on typographical, non-word spelling errors.

In most cases, and certainly in this case for te reo Māori, spelling error detection involves matching text to a dictionary or list of words. The dictionary is more usually just a list of correct words in the target language, however, it would be reasonable and sensible to expect the inclusion of some popular loan words from one or more other languages – a New Zealand English dictionary would also contain a number of te reo Māori words for example. Each word in a given text is then matched to the dictionary and returned as an error if that word, as spelt, is not contained in the dictionary.[vi] There are drawbacks to this method which include the requirement to ensure such a dictionary is kept current and sufficiently extensive to cover all the words in a text[vii] and to include commonly used loan words from other languages, *kia ora, kiwi,* or *Aotearoa* in NZ English for instance.

Once an error is detected, the spellchecker assumes that the misspelled word is a mutation of one (or more) of the words contained in the dictionary. Correction then consists of two steps:

1. generating candidate corrections, and,
2. ranking those corrections.

Generating candidate corrections usually makes use of a precompiled table of legal n-grams to provide one or more potentially correction replacement – where n-grams generally refer to the number of contiguous *n* letters within a word. Ranking the candidate corrections is a process that identifies some measure of lexical similarity between the misspelled word and the suggested corrections, or a probability-based estimate of the likelihood that the suggested replacement is most accurate in order to rank the candidate corrections. These two parts of spelling correction algorithms are usually treated as separate processes and executed in sequence but, at times, the ranking and final selection is a manual, subjective choice performed by the user.[vii] In short, spelling correction algorithms will calculate the amount of changes required to provide correct words. The amount of changes is referred to as the 'distance' (between the misspelled word and each word in the dictionary) which provides a similarity and likelihood score that is then used to rank the corrections or replacement suggestions based on how short that distance is.[vi]

A more technical overview with diagram is provided in Appendix 2: Spellchecking – Basic Technical Overview.

5 Existing Providers

Initial investigations have identified current initiatives with some level of relevance in terms of spellchecking, translation, and learning and teaching te reo Māori[2]:

- Te Ngutu Kura
    - Free to download spell checker
    - Can be used with most major platforms and operating systems
    - Has an option to spell check words with double vowels
    - Word list is freely available to developers
    - More than 58,000 word list entries including NZ place names
- Stars21
    - Provides spellchecking for Te Reo Māori
    - Also an online translation tool powered by Google Translate
- Wakareo ā-ipurangi
    - Owned by Wordstream
    - Te Reo Tupu – Māori-English-Māori compilation dictionary
    - Searches are browser based
    - Spellchecker (Moana Kupu Whaiutu) not found
- Te Aka – Māori Dictionary
    - Online bilingual dictionary
    - Includes encyclopaedic entries, idioms and grammatical explanations
    - Indexed to the Te Whanake Māori language series
- Ngata Dictionary
    - English-to-Māori, Māori-to-English with sentence examples
    - A wide range of contemporary and traditional contexts
- He Pātaka Kupu
    - Te Reo Māori dictionary (Te Taura Whiri i te reo Māori)
- Kupu
    - In conjunction with Te Taura Whiri i te Reo
    - Kupu o te rā/wiki
    - Includes word lists, tests, parts of speech and sentence structures
- Te Kupenga Hao i te Reo
    - Web-based resources
    - Includes Paekupu – contemporary curriculum-based word lists
- Te Kete Ipurangi
    - Curriculum-based resources
    - Includes Mātaiko, Reomations, He Kohinga Rauemi ā-Ipurangi
- Kauwhata Reo
    - Online resources to support learning language and some aspects of tikanga
- Te Whanake
    - Māori language learning online
    - Based on the Te Whanake textbooks
    - Includes videos, apps and podcasts
- Whakairo Kupu
    - Quizzes for learning, reading and writing te reo
    - Includes a flashcard section
- Kotahi Mano Kaikā

[2] Links to these resources have been supplied in Appendix 1: Links to existing providers

    - Ngāi Tahu initiative to have 1,000 homes speaking te reo Māori
- New Zealand History
    - 100 words every New Zealander should know
    - 1,000 place names
    - 365 useful words and phrases in te reo Māori
- Māori Language.net
    - Resources for learning and practising te Reo
    - Includes video lessons, podcasts, songs and phrases

Of the identified spellcheckers, one (Te Ngutu Kura) uses a dedicated word list, another (Stars21) is browser-based powered by Google Translate, and the third (Moana Kupu Whaiutu from Wakareo ā-ipurangi) was not found – the link to the web page was broken. We have discussed the acquisition and incorporation of the Te Ngutu Kura word list because it contains over 58,000 words that include NZ place names. However, we are unsure when this corpus was created, how it was created and by who. Given that the process we employed had some success this list could be reconsidered. Manually vetting such a large list as the final quality check may have to be reviewed and somehow automated.

6 NZQA Considerations- how can it work?

The SoNET software matches words in the target text to words stored within an English (Australian) dictionary that was master-sourced online. The word matching is actioned letter by letter. The correct word suggestions in the event of misspelled words are also derived from the master dictionary and actioned using a 3rd Party Plug-in (TinyMCE). The current English language dictionary numbers 49,437 words (May 2020).

Feedback (and intuition) suggests accuracy would be better with more words in the dictionary. However, we are unsure how the larger word lists would affect processing times for the different users who may have varying levels of online capability.

In terms of the minimum dictionary size for te reo Māori, we consider a sample size of approximately 15,000 kupu would be sufficient for proof of concept testing and quite ample for a pilot program. Proper nouns, the names of places and people, are treated the same way as other words in the dictionary. The software checks for correct letter sequence, correct case and supports UTF8 – which means the system is able to use and check for macrons.

Currently, the spell check application is linked to one dictionary and is user-activated i.e. the checker does not function while the user is typing. We consider that the action of the current spellcheck software will not be affected by Māori words since the letters used are the same as for English and the only difference is the use of macrons for te reo Māori. With this in mind, more importance should be placed on ensuring the format of the Māori dictionary is either the same as the English one or similar enough so that it can be used by the spellchecking software without issue. Our queries regarding the occurrence of suffixes against some of the words in the English dictionary were not fully answered by the SoNET advisors and we assumed they were grammar markers – parts of speech identifiers such as verbs and nouns. Given that many of the words in the currently-used dictionary were without these markers we assumed that the suffixes would not be a critical feature of the file format for spellchecking te reo Māori. Our strategy therefore would include the creation of a Māori word dictionary to be used by the spellchecking software as part of a controlled test. The aim of such a test would be to determine whether the new dictionary, which potentially included kupu Māori, works with the software and to test how accurately the software identifies incorrect te reo Māori text.

Testing Options:

1. to develop one large inclusive dictionary by adding Māori words to the dictionary already in use. The spellcheck software would then highlight words that didn't match anything in the list but potentially only be able to provide replacement suggestions for the words that were part of the original English dictionary. This would also satisfy, at least in part, a query from NZQA about words that were not English but were used in everyday conversation – Aotearoa was mentioned but other examples include kia ora, kiwi, mihi, haere, Māori and Māori placenames;
2. to provide a separate dictionary for te reo and develop a software enhancement that allows the spell checker to switch between dictionaries understanding that there will be some cross-over (loan) words.

Option 1 would seem to be the preferred option given the occurrence of loan words, instances of everyday bilingual speech and the use of reo Māori proper nouns – names of people and places for instance. There appears to be no immediate requirement to modify the existing software to accurately apply spellchecking for te reo Māori.

While we expect the proof of concept testing will support spellchecking for te reo Māori in the first instance, discussion and thought should be applied to dialectic and regional differences of words and language use. For instance, students from Waikato may want to use tētehi for tētahi or kōrerongia for kōrerotia. Students from te Wai Pounamu may want to switch ng for k, so rangi becomes raki. Additionally, the passive forms of many verbs include more than one usable suffix. These sorts of issues will need to be addressed when developing the te reo Māori dictionary that will be used within the working system.

Word replacement suggestions for te reo Māori would either require a deeper investigation into how the 3rd party plug-in (TinyMCE) works, a wider search for an add-on that works for te reo or can be effortlessly repurposed, or the development of a custom-designed tool.

7 Rārangi Kupu

Given high quality word lists are crucial for accurate spellchecking, considerable effort was spent creating an accurate word list. A range of corpora of Māori-language words were identified and acquired through the appropriate channels for each source.

Corpus	No. of Words
1. Paekupu[3] covering the following curriculum areas:	
a. Te Reo Matatini (*Literacy*)	1103
b. Toi Ataata (*Visual Arts*)	1149
c. Puoro (*Music*)	606
d. Ngā Mahi a Rēhia (*Recreation*)	912
e. Pāngarau (*Mathematics*)	1194
f. Pūtaio (*Science*)	2309
g. Hangarau (*Technology*)	1117
2. PapaKupu03[4]	21,916
3. PapaKupu04[4]	23,097
4. Williams Dictionary (WilliamsDictWords_POS_v0.1,)	13,598
5. Reo Hangarau[5]	465
6. Kuputaka	3,528
7. Hangarau o te Ao[6]	924

A single rārangi kupu, te reo Māori word corpus numbering 71,930 words, was created from the above corpora. We had no expectation regarding an upper word count limit but assumed that more words would be better although not as important as word accuracy. Ensuring the words contained in the list were valid meant that the provided list needed to be vetted by a language expert so it would be of as high a quality as possible. We briefly considered lower word list limits but only in the context of a smaller sample word list suitable for proof of concept controlled testing. This sample list was completed and sent to NZQA in preparation for proof of concept testing mid-October 2020.

The above corpora were provided in a spreadsheet (.xls) file format and were transferred to a text (.txt) format for simplified filtering and sorting using a customised program. The programming language Java was used so that the code could easily be based on some already established methods in the Java library. The program was designed to:

1. Remove English words
    a. by detecting letters that are not in the Māori language
    b. by detecting words ending in consonants.
2. Remove other foreign characters (non-letters)
    a. Using a regex (regular expression) condition e.g. ":|,|;|=|\\.|\\(|\\)|\\$|/".
3. Delete white spaces
    a. words following a white space were split and placed on a separate line
    b. repeated for all white spaces and multiple words on a line.

[3] Provided by Ian Christensen of Te Kupenga Hao i te Reo - 2020
[4] Sourced by Associate Professor Te Taka Keegan (University of Waikato) - 2017
[5] Provided by Te Mihinga Komene – he kohinga kupu iPae, matihiko hoki - 2020
[6] Sourced from Tā Ian Cormack - 2020

4. Sort the list using inbuilt Java functionality:
    a. alphabetically,
    b. remove all duplicate words
    c. convert all letters to lower case.

Extracting English words and words with foreign characters resulted in the removal of 21,108 words. A further 30,409 words were removed once the white spaces and duplicate words had been accounted for. The list was then put through macronising software which automatically places the macron in the appropriate places. The final working list comprised 16,569 words.

Removing duplicate words proved remarkably similar to spellchecking. Using a basic word/letter comparison technique, each word was compared letter by letter with the word that followed in the alphabetically sorted list. The letter-by-letter comparison technique proved quite efficient. As soon as the test returned a false condition, meaning one of the letters did not match, the program assigned uniqueness and immediately moved to the next word. When the letter-by-letter comparisons reach the end of the word and all letters match, a true condition was returned, and the duplicate word was deleted.

All removed words and foreign characters were stored in separate lists to enable later feedback that might help to improve the quality of each of the original lists and to make sure the program algorithms were performing the correct checks and performing them correctly. The program was re-run over the final list and then the list was manually checked for any obvious errors. Some errors were identified that sat out of the parameters of the algorithms. These were rectified and the list was again rechecked for other obvious errors.

A registered interpreter/translator was employed to validate the accuracy and quality of the word list. Non-Māori words, words that should have been capitalised and words that required macrons to be added or deleted were identified and flagged. The flagged words have been removed or actioned as required and the final vetted list has been provided separately. The final vetted list comprises 14,132 words.

8 Further Considerations

The use of a dedicated dictionary for word matching necessitates a regular cycle of currency testing. As words evolve or appear, generated by new subject areas, practices or technologies, they should be vetted and included in the dictionary to ensure the input text from students is checked against a list that is as current and as relevant as possible. We expect that Te Kupenga Hao i te Reo (and Paekupu) in conjunction with the Ministry of Education would cover word currency in terms of contemporary curriculum. Therefore, we suggest a process of vetting that results in an updated list every five years (at most).

As previously mentioned, the accuracy of the spellchecking process is largely dependent on the quality of the dictionary and the accuracy of the words contained within. Dialectic and regional differences and grammatical traits such as multiple verb-ending possibilities should be considered and the variants added to the word list.

It seems that the nature of current checking (letter-by-letter) is cumbersome. A method that looks at exact word matching initially, that is then followed by a letter-by-letter check when a potential misspelled work is detected, would be much more efficient. Other methods of word checking could also bear investigation.

Word suggestions could also be derived from:

- Matching part words
- Probability accuracy (based on letters within the words)
- Dynamic programming algorithms – efficiency costing
- Word prediction using bi-grams and/or tri-grams
- Checking based on keyboard proximity.

Other methods of automatic spell checking not discussed here that might involve development, customisation or software acquisition are more likely a part of a larger strategic conversation.

9 Summary and Conclusions

Ideally, the capability of spellchecking for te reo Māori would be dependent on the accuracy of relevant data and the application that is then deployed to use that data. We have sought to test the accrued corpus with the SoNET software but timing and technical considerations have thwarted early testing bids. In terms of a deployable word list format, a sample word list was provided for proof-of concept testing, followed by the supply of the full version 14,132 words. Initial queries regarding design requirements of existing software culminated in subsequent advice to reformat the word list so that is could be used by the current software. The modifications were actioned in accordance with NZQA technical staff guidelines. The marked-up file has since been made available but has yet to be vetted or deployed.[7]

In its simplest form, spellchecking can be split into a mechanical procedure of matching input words against a dictionary followed by a more statistical process that suggests valid word replacements for any error that is detected. In practice, how the word is tested is of less importance when compared to how accurately it matches that word to one that is contained in the dictionary. Whether the word is tested in its entirety or matched letter-by-letter is relatively irrelevant in the context of the returned result. For example, basic English spellchecking does not recognise instances of legitimate words that have been miss-used in particular contexts. Furthermore, most spellcheckers don't recognise colloquialisms or regional slang. In terms of te reo Māori, the difficulties of accuracy are similar but would also revolve around dialectal differences and colloquiality. This then means that the dictionary for te reo Māori that is used for first-pass spellcheck would need to be extended if it were to be used in educational digital assessment of people from a variety of backgrounds. We expect that the ongoing development of the dictionary would require tribal and regional dialectal support.

In terms of this particular research, there are a variety of initiatives that support spellcheck capability for te reo Māori. In terms of online educational assessment, that functionality requires much work but, in context, the amalgamation of appropriate data and technology is not insurmountable. An understanding of the technology in use, including the functionality of the third-party plugin and the process of correction suggestion for word errors, would be useful. However, we submit that this research has provided a starting point, where the data is accurate for proof-of-concept and potentially usable by the application in use.

[7] At time of writing - while acknowledging short timeframes

10 Appendices

10.1 Appendix 1: Links to existing providers

Te Ngutu Kura	https://www.taiuru.maori.nz/publications/digital-tools/te-ngutu-kura/
Stars21	http://www.stars21.com
Te Aka	https://maoridictionary.co.nz
Ngata Dictionary	https://www.teaching.co.nz/dictionary
He Pātaka Kupu	https://hepatakakupu.nz
Wakareo-ā-ipurangi	http://www.reotupu.co.nz
Te Kupenga Hao i te Reo	https://kupengahao.co.nz/web-resources/
Paekupu	https://paekupu.co.nz/#
Kupu	https://kupu.maori.nz/kupu
Whakairo Kupu	https://quizlet.com/310903417/whakairo-kupu-flash-cards/
Te Kete Ipurangi	https://www.tki.org.nz
Kauwhata Reo	https://kauwhatareo.govt.nz/
Te Whanake	https://tewhanake.maori.nz
Kotahi Mano Kaikā	http://www.kmk.maori.nz
NZ History - 100 words	https://nzhistory.govt.nz/culture/maori-language-week/100-maori-words
NZ History - 1,000 words	https://nzhistory.govt.nz/culture/maori-language-week/1000-maori-place-names
NZ History - 365 phrases	https://nzhistory.govt.nz/culture/maori-language-week/365-maori-words
Māori Language.net	https://www.maorilanguage.net

10.2 Appendix 2: Spellchecking – Basic Technical Overview

We can use the equation P(c|w) for which the probability (P) of a user-given word '*w*', the spell checkers job would be to find and produce the correct spelling '*c*' of '*w*'.

Given a user input '*w*', there is no way to 100% know its correct spelling '*c*' (for example, should "nights" be "knights" or "night"). A good way to model the spelling check problem is to use a probabilistic approach; we are trying to find the correction '*c*', out of all possible candidate corrections 'C', that maximizes the probability that '*c*' is the intended correction, given the original word '*w*':

$$argmaxc \in C\ P(c|w)$$

By Bayes' Theorem this is equivalent to:

$$argmaxc \in C\ P(c)\ P(w|c) / P(w)$$

Since P(w) is the same for every possible candidate '*c*', we can factor it out, giving:

$$argmaxc \in C\ P(c)\ P(w|c)$$

The four parts of this expression are:

Selection Mechanism: argmax

We choose the candidate with the highest combined probability.

Candidate Model: $c \in$ candidates ‘C’

This tells us which candidate corrections, c, to consider. When generating the candidate set ‘C’, we can make use of the Edit Distance metric. A rule of thumb of spelling check is that 75% misspells are within edit distance 1 and 98% are within edit distance 2. Thus in this project, we will limit the candidate set ‘C’ only to those words that are within edit distance 2 of the user provided input w.

Language Model: P(c)

The probability that c appears as a word of English text. For example, occurrences of "the" make up about 7% of English text, so we should have P(the) = 0.07.

Error Model: P(w|c)

The probability that w would be typed by a user when the user meant c. For example, P(teh|the) is relatively high, but P(theeexyz|the) would be very low.

The ‘probabilistic approach’ doesn’t have a very high accuracy but is one of the simpler methods of a spell checker. Accuracy from spell checker tests range from 68%-75%.[8]

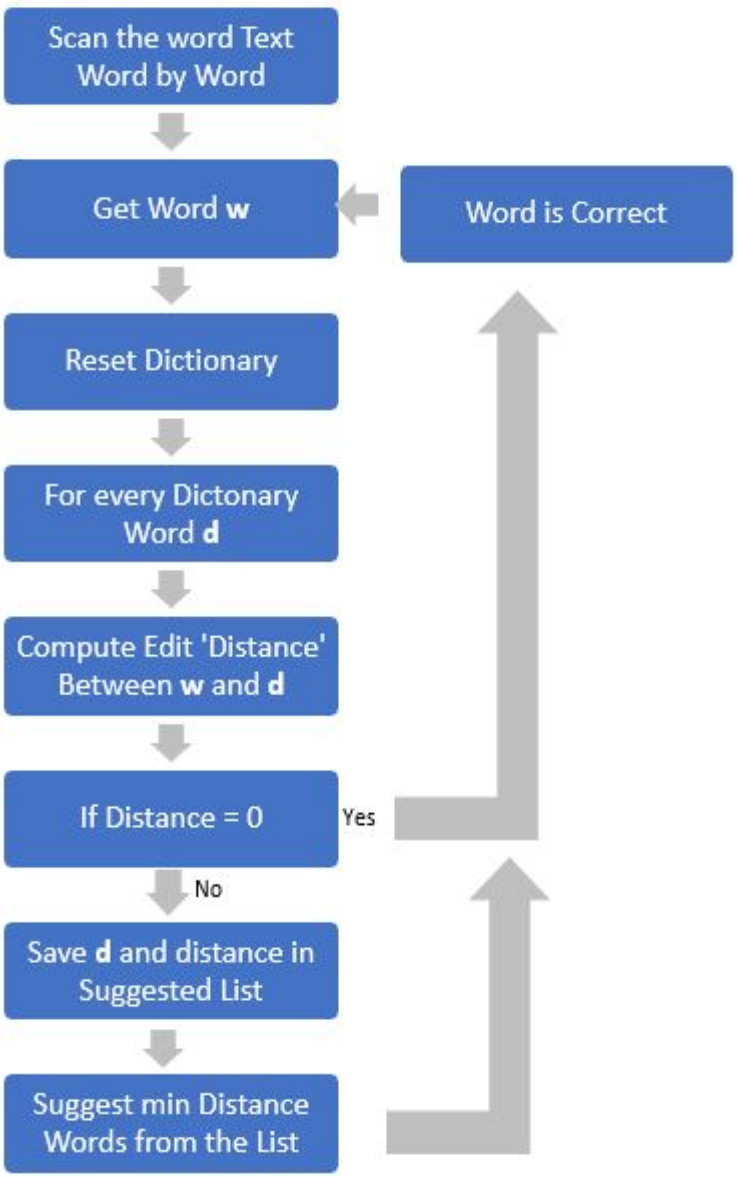
Spell Checking in Practice

This diagram is an interpretation of the basic spellcheck process:

[8] See: *http://norvig.com/spell-correct.html*

Scan word:

The scanned word is stored (w) and held for comparison.


Scan the word Text Word by Word
Get Word **w**
Word is Correct
Reset Dictionary
For every Dictonary Word **d**
Compute Edit 'Distance' Between **w** and **d**
If Distance = 0
Yes
No
Save **d** and distance in Suggested List
Suggest min Distance Words from the List

Load or reload the dictionary file for cross-referencing.

Each word is stored one-at-a-time (d) and compared to (w).

(w) is scanned letter-by-letter. Computing (matching) can be modified.

Distance between 'w' and 'd' is computed.

This returns an integer 0 if the characters are the same and a 1 if they are not.

Compare the characters from each word 'w' and 'd'.

Comparisons are made by matching first character 'c' of each word and comparing the two. Then matching the second set of characters in each word and so on.

11 EndNotes

[i] Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental 5*(1). Retrieved from https://cs.slu.edu/~scannell/pub/wac3.pdf

[ii] Gordon, R. (Ed.) (2005). *Ethnologue: Languages of the world* (15th ed.). Dallas, Texas, USA: SIL International.

[iii] UNESCO. (2009) *UNESCO Atlas of the World's Languages in Danger.* Retrieved from http://www.unesco.org/culture/languages-atlas/index.php

[iv] Whitelaw, C., Hutchinson, B, Chung, G.Y., & Ellis, G. (2009). Using the Web for Language Independant Spellchecking and Autocorrection. Retrieved from https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/36180.pdf

[v] Kumar, R., Bala, M., & Sourabh, K. (2018). A study of spell checking techniques for Indian Languages. *JK Research Journal in Mathematics and Computer Science,* 1(1), 105-113. Retrieved from http://jkhighereducation.nic.in/jkrjmcs/issue1/15.pdf

[vi] Engineer by Day. (2012). How Spell Checkers Work – Part 1. Retrieved from https://engineerbyday.wordpress.com/2012/01/30/how-spell-checkers-work-part-1/

[vii] Etoori, P., Chinnakotla, M., & Mamidi, R. (2018). Automatic Spelling Correction for Resource-Scarce Languages using Deep Learning. Retrieved from https://www.aclweb.org/anthology/P18-3021.pdf