RESULTS

Version 2
27/04/2020

TABLE OF CONTENTS

01

INTRODUCTION

Scores and Datasets Explained

02

METHODOLOGY

Method set up

03

RESULTS

Scores, Overlap and Interesting Word Corrections

01

INTRODUCTION

Methods and Datasets

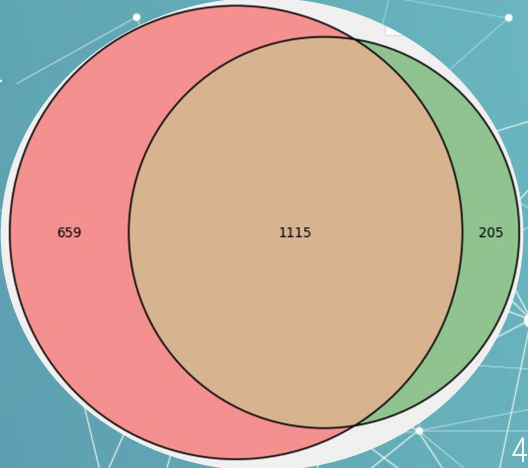
INTRODUCTION

This is an overview of the results generated from the work on the phonetic acoustic edit distance spelling correction method. In the results there are several scores which are displayed:

- **Accuracy** - the percentage of correct word corrections (using the suggested word to correct with)
- **Correct in candidates** - If correct word not suggested, percentage of times correct word was in candidates
- **Recall** - A combination of the accuracy scores and the candidates scores. The percentage of times the correct word was suggested or was contained in the candidates list.

There is also an overlap graphic of each traditional spelling tool versus the phoneme approach. The values shown in these graphics are as follows:

- **Red** - denotes unique corrections method by the chosen traditional method
- **Brown** - denotes corrections made by both the traditional and phoneme method
- **Green** - denotes unique corrections made only by the phoneme method



SPELL CHECKING METHODS

Currently there are four spell checking tools used in this study:

PySpellChecker

Generates all possible terms for a word within 2 edit distance (deletes + transposes + replaces + inserts) from the query term and then searches in the dictionary.

SymSpell

Generates terms with an edit distance of 3 (deletes only) from the dictionary, and then adds these terms along with the original term to the dictionary.

Aspell Approach

Uses the GNU Aspell open source spelling correction tool.

Phoneme Approach

A fork of SymSpell but altered for phoneme sequences. Using an acoustic distance matrix to calculate the edit distance between phoneme sequences.

CORPORA OF MISSPELLINGS USED PUBLIC

For testing and comparison, publicly available corpora of misspellings from Birkbeck University of London were used:

- **Birkbeck** – 36,133 misspellings of 6,136 words. Errors amalgamated from native-speaker section of the Birkbeck spelling corpus. Includes results of spelling tests and errors from free writing taken mostly from schoolchildren, university students or adult literacy students.
- **Holbrook** – 1,791 misspellings of 1,200 words. Extracts of writings of secondary-school children in their penultimate year of school.
- **Aspell** – 531 misspellings of 450 words. Used for testing GNU Aspell spell checker.
- **Wikipedia** – 2,455 misspellings of 1,922 words. List of misspellings made by Wikipedia editors, found here.

CORPORA OF MISSPELLINGS USED
PRIVATE

Data provided by Zeeko, a bullying education company in Nova UCD:

- **Zeeko Dataset** – 232 misspellings of 163 words. Gathered from fifteen Zeeko surveys carried out by school children in Ireland. Free-text field input on a submitted survey. Due to it being submitted via computer the dataset may be more susceptible to typos (keyboard strokes) or auto corrects.
- Misspellings were hand labelled by referencing the context of the misspelling and interpreting the presumed correct spelling.
- Where a judgment could not be made, the misspelling was excluded from the dataset.

02

METHODOLOGY

METHODOLOGY

PySpell - uses two edit distances for suggestion and candidate list generation. Uses its own dictionary.

SymSpell - uses two edit distances for suggestion and candidate list generation. Uses its own frequency dictionary.

Aspell - installed and run as is. Uses its own English language package dictionary.

Phoneme Method - two edit distances in a phoneme sequence. Acoustic edit distance insert/delete cost = 1.0. Uses the CMU dictionary.

03

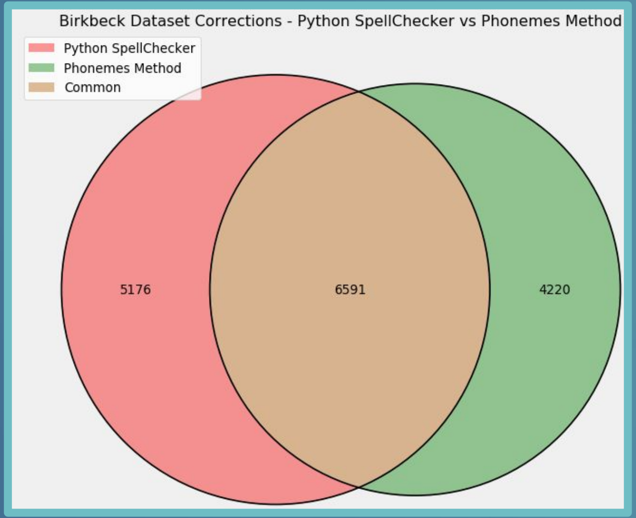
RESULTS

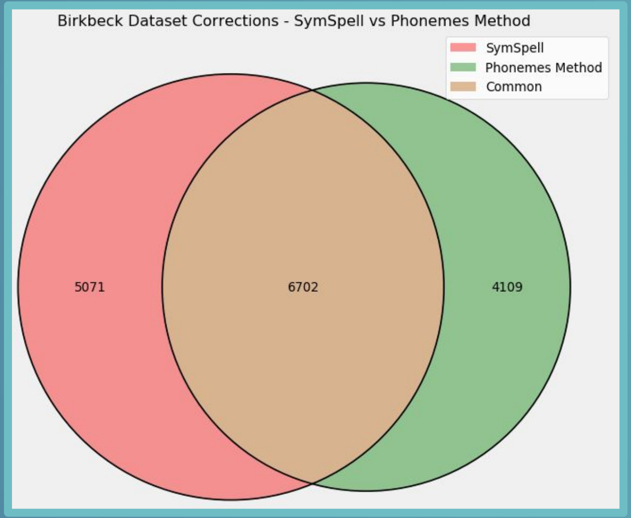
Scores, Overlap and Interesting Word Corrections

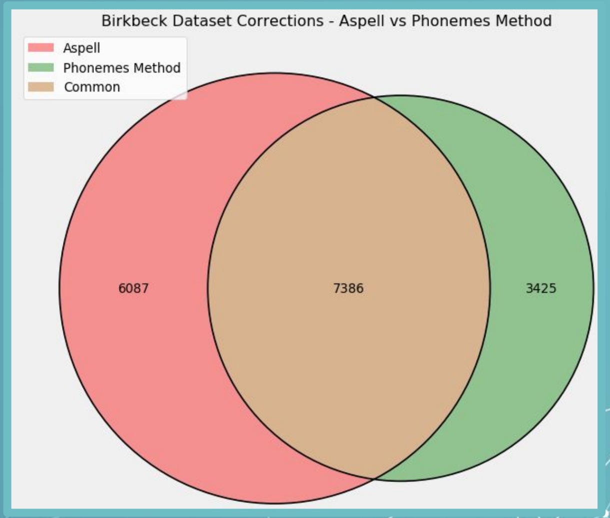
BIRKBECK SCORES

	Accuracy	Correct in Candidates	Recall (Correct + Cand)
PySpell	34.72%	7.03%	41.75%
SymSpell	34.74%	8.30%	43.04%
Aspell	39.76%	25.57%	65.33%
Phonemes Method	31.90%	15.20%	47.10%

BIRKBECK OVERLAP



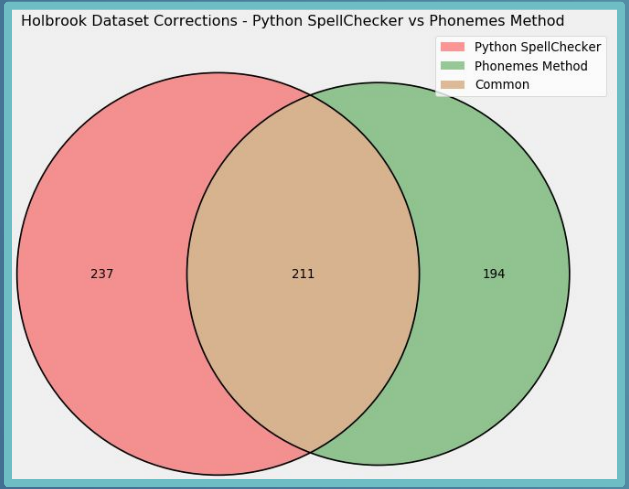


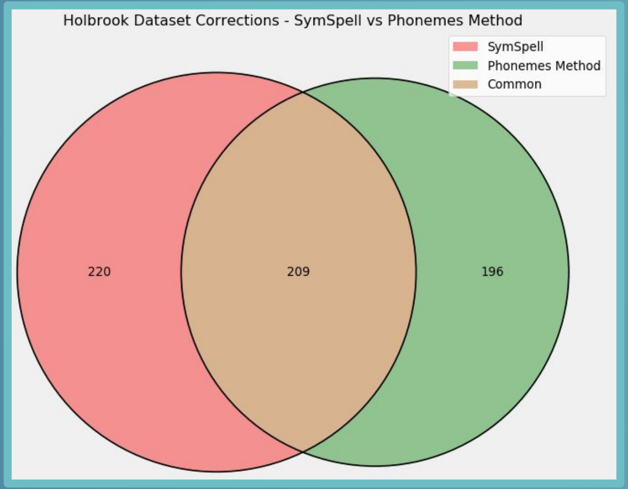


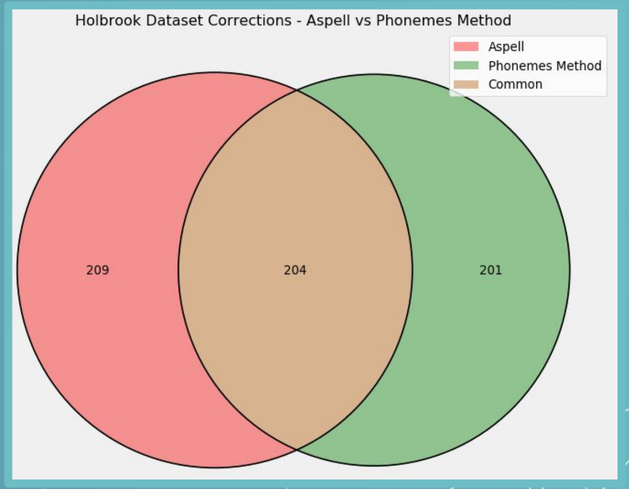
HOLBROOK SCORES

	Accuracy	Correct in Candidates	Recall (Correct + Cand)
PySpell	28.68%	11.84%	40.52%
SymSpell	27.46%	15.04%	42.51%
Aspell	26.44%	37.96%	64.40%
Phonemes Method	25.93%	19.85%	45.77%

HOLBROOK OVERLAP



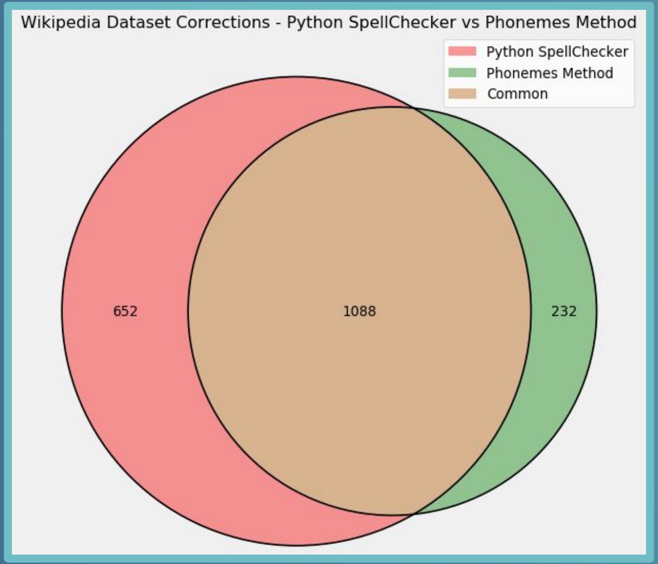


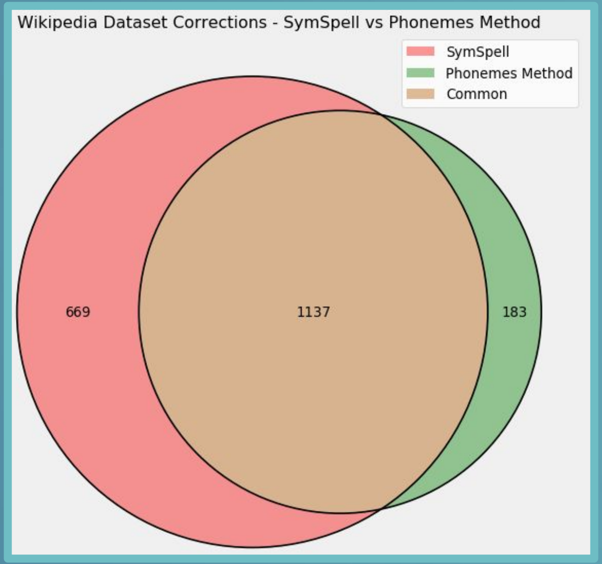


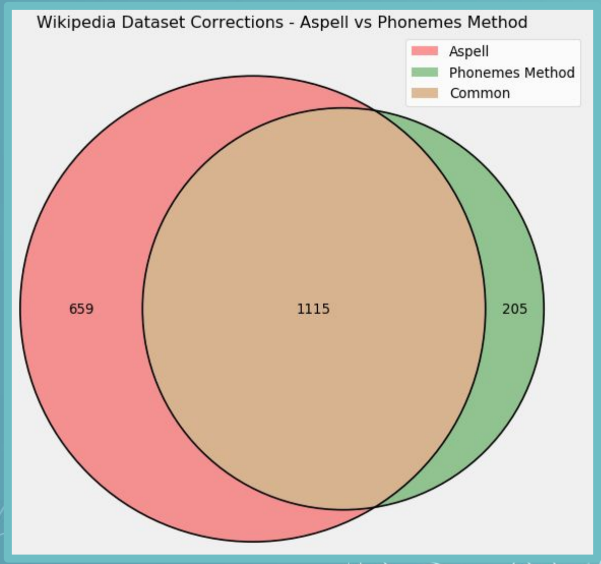
WIKIPEDIA SCORES

	Accuracy	Correct in Candidates	Recall (Correct + Cand)
PySpell	78.03%	9.69%	87.71%
SymSpell	80.99%	11.12%	92.11%
Aspell	79.55%	13.63%	93.18%
Phonemes Method	59.19%	12.15%	71.35%

WIKIPEDIA OVERLAP


Wikipedia Dataset Corrections - Python SpellChecker vs Phonemes Method

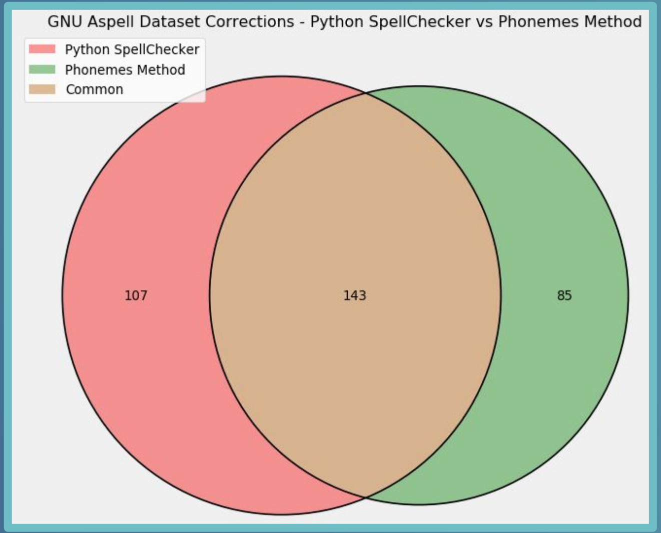

Wikipedia Dataset Corrections - SymSpell vs Phonemes Method

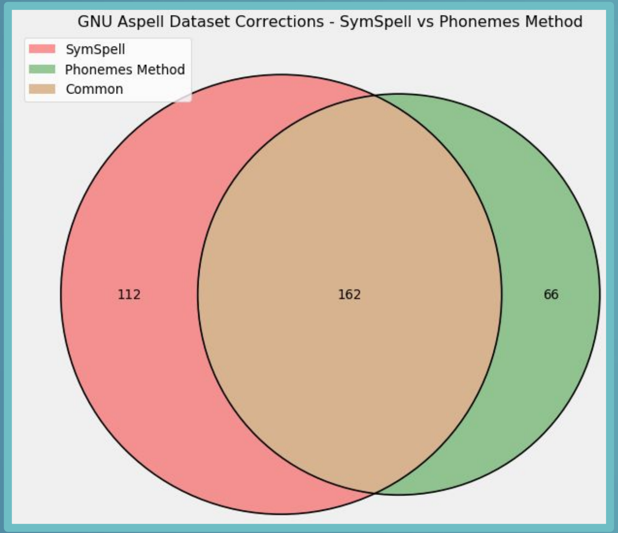

Wikipedia Dataset Corrections - Aspell vs Phonemes Method

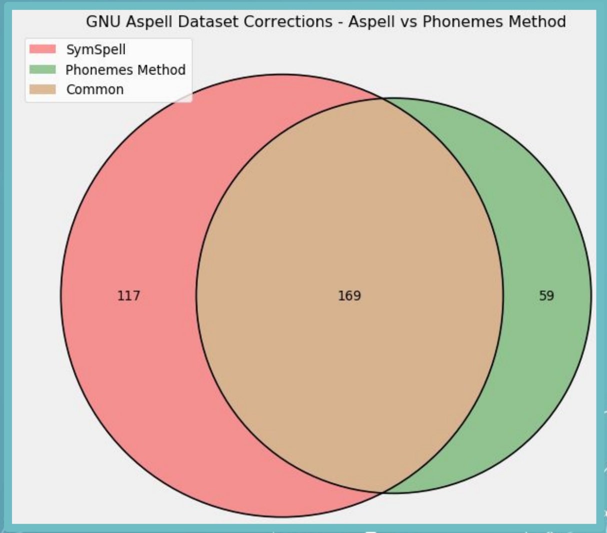
ASPELL SCORES

	Accuracy	Correct in Candidates	Recall (Correct + Cand)
PySpell	48.54%	12.62%	61.17%
SymSpell	53.20%	13.98%	67.18%
Aspell	55.53%	30.10%	85.63%
Phonemes Method	44.27%	16.50%	60.78%

ASPELL OVERLAP


GNU Aspell Dataset Corrections - Python SpellChecker vs Phonemes Method

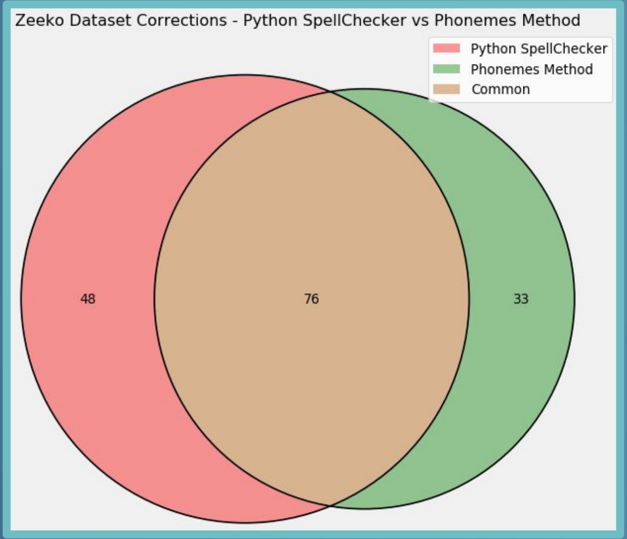

GNU Aspell Dataset Corrections - SymSpell vs Phonemes Method

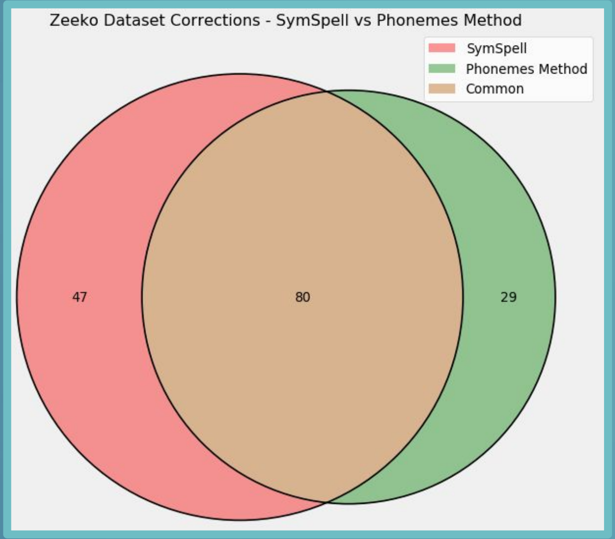

GNU Aspell Dataset Corrections - Aspell vs Phonemes Method

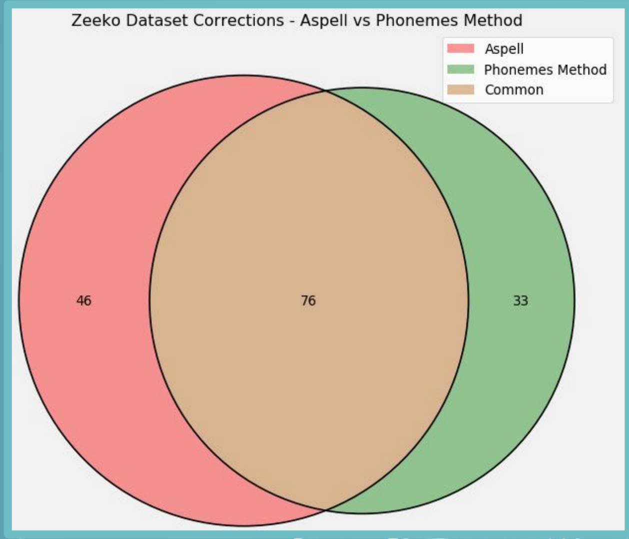
ZEEKO SCORES

	Accuracy	Correct in Candidates	Recall (Correct + Cand)
PySpell	53.45%	12.50%	65.95%
SymSpell	54.74%	15.95%	70.69%
Aspell	52.59%	26.72%	79.31%
Phonemes Method	46.98%	21.12%	68.10%

ZEEKO OVERLAP







INTERESTING WORD CORRECTIONS

	Misspelling	Phoneme Rep	Suggested Phoneme Rep
educational	egicasinol	EH JH AH K EY S IH N AO L	EH JH AH K EY SH AH N AH L
example	egsample	EH G S AE M P AH L	IH G Z AE M P AH L
situation	sichweshen	S IH CH W EH SH AH N	S IH CH UW EY SH AH N
enjoyed	injoid	IH N JH OY D	EH N JH OY D
destroyed	distroid	D IH S T R AA D	D IH S T R OY D