A Large List of Confusion Sets for Spellchecking Assessed Against a Corpus of Real-word Errors

Jennifer Pedler, Roger Mitton
Birkbeck, University of London
Malet Street, London WC1E 7HX, UK
j.pedler@dcs.bbk.ac.uk, r.mitton@dcs.bbk.ac.uk

Abstract

One of the methods that has been proposed for dealing with real-word errors (errors that occur when a correctly spelled word is substituted for the one intended) is the "confusion-set" approach - a confusion set being a small group of words that are likely to be confused with one another. Using a list of confusion sets drawn up in advance, a spellchecker, on finding one of these words in a text, can assess whether one of the other members of its set would be a better fit and, if it appears to be so, propose that word as a correction. Much of the research using this approach has suffered from two weaknesses. The first is the small number of confusion sets used. The second is that systems have largely been tested on artificial errors. In this paper we address these two weaknesses. We describe the creation of a realistically sized list of confusion sets, then the assembling of a corpus of real-word errors, and then we assess the potential of that list in relation to that corpus.

A "real-word error" occurs when a correctly spelled word is substituted for the one intended, as in, "The Wine Bar Company is opening a chain of *brassieres*." These account for perhaps a quarter to a third of all spelling errors (Mitton, 1996) – *there* for *their*, *principle* for *principal* and the like – and the problem may have been exacerbated by the widespread use of spellcheckers; perhaps the writer of the above wrote *braseries*, the spellchecker proposed *brassieres* as the first on its list and the writer, without paying much attention, accepted it. (There is even a name for these errors – "Cupertinos". The term arose because an earlier version of Microsoft Word would query *cooperation*, having only the hyphenated *co-operation* in its dictionary, and would offer *Cupertino*, which is the name of a suburban city in California, as its first suggestion; this has given rise to some official documents containing phrases such as "agreement on bilateral Cupertino".)

One of the methods that has been proposed for dealing with real-word errors is the "confusion-set" approach. In the classic version, a list of confusion sets is drawn up in advance of the spellchecking, a confusion set being a small group of words that are likely to be confused with one another, such as *principle* and *principal*. When checking a text, the spellchecker looks out for any of these words. If it finds one (say *principal*), it retrieves the other words from that word's confusion set (here just *principle*) and assesses whether one of these other words would be a better fit at that place in the text; this could be on the basis of syntax, semantics, probability, some combination of these or any other information that could be brought to bear. If one of these other words in the confusion set appeared to be a better fit, this word would be proposed as a correction.

Confusion sets are not confined to homophones; if writers occasionally wrote *hopping* for *hoping* or *minuets* for *minutes*, then these would be candidates for a confusion set. Nor is the method confined to correcting mistakes of spelling; it can just as well be applied to errors of usage, such as the confusion of *between* and *among*.

Much of the research on the confusion-set approach has suffered from two weaknesses. The first is the small number of confusion sets. Researchers (e.g. Golding, 1995; Golding and Roth, 1999; Jones and Martin 1997, Golding and Schabes 1996, Carlson and Fette, 2007) have typically used lists of about 20 sets (in fact often the same list, to preserve comparability with earlier work); to have any hope of real-life applicability, the list would have to run well into the thousands. The second is that they have tested their systems on artificial errors. While adequate for proof of concept, this leaves open the question of how well the systems would perform in spellchecking actual text.

Researchers have not been unaware of these weaknesses and some efforts have been made to deal with them, or at least with the first of them. Carlson et al. (2001) experimented with a list of 265 confusion sets – still well short of a real-life number but an improvement on 20. Another approach, used in early work by Mays et al. (1991) and more recently by Fossati and Di Eugenio (2008) is to dispense with a predefined list of confusable words and to generate alternatives for most of the words in the text. For each word in the text being spellchecked, Mays et al. created a confusion set on the fly by extracting from their dictionary all the words (typically two to five) that differed from it by a single-letter edit. Fossati and Di Eugenio did not use a dictionary but derived a vocabulary of about 9000 words from a training corpus and precomputed for every word in the vocabulary a set of other words from the same vocabulary that resembled it orthographically or phonetically; these sets contained, on average, 86 words and the largest had 445, far exceeding the two or three words in the confusion sets generally used in earlier work (figures kindly supplied by Dr Fossati).

As to the second weakness, however, all researchers seem to have reluctantly accepted it simply because collections of genuine real-word errors, in context, are hard to come by. Either they have fallen back on test corpora of artificial errors, i.e. taking a piece of correct text and simply changing some of the occurrences of *principle* to *principal* and so on, or they have tested their systems on correct text. The rationale for this latter approach is that, if the text is correct, the system should not propose any corrections; if, say, the text contains a correct occurrence of *principle* and the system, after considering whether to change it to *principal*, decides to leave it as *principle*, then it has made a correct decision. There is some sense in this but, even so, it seems peculiar to test a text-correction system on text that does not need correcting.

In this paper we address these two weaknesses. We describe the creation of a realistically sized list of confusion sets, then the assembling of a corpus of real-word errors, and then we assess the potential (or, if you prefer, the limitations) of that list in relation to that corpus.

1.A Realistically-sized List of Confusion Sets

To create a realistically large list of confusion sets, we needed to consider the types of error users are likely to make. We began with the spellchecker developed by Mitton (1996), which ranks its suggestion lists using a version of the well-known string-to-string edit-distance algorithm (Levenshtein, 1966; Wagner and Fischer, 1974; Veronis 1988). This assigns a cost to each single-letter insertion, deletion or substitution required to transform one string into another; the lower the total cost, the more similar the strings. The system has been tuned, using a large collection of predominantly non-word errors, to assign a lower score to the type of mistakes that users are more likely to make. For instance, inserting the missing *c* in *sissors (scissors)* would have a lower cost than it would in *satter (scatter)* on the grounds that people are more likely to omit the *c* from *scissors* than the *c* from *scatter*. To produce an initial list of possible confusables, we ran this program over the dictionary, comparing each word with every other word and storing the pairs that scored less than a predefined threshold.

The resulting list contained just over six thousand pairs of words. These were in the form of <*a,b*> word pairs with each pair listed once; thus, for example, the list included the pair <*bad, bade*> but not the pair <*bade, bad*>. Although the pairs were unique, each individual word could occur more than once either as word *a* or as word *b*. *Bad*, for example, appears five times as a word *a*; it is also paired with *bard, bawd, bed* and *bid*; *write* is a word *b* in the pair <*writ, write*> and a word *a* in the pair <*write, writhe*>. The order in which the words appear in these pairs and whether they appear as word *a* or word *b* is simply a function of the ordering of the words in the dictionary.

A number of pairs in this initial list were unsuitable for inclusion in confusion sets – proper nouns, prefixes, abbreviations and variant spellings (e.g. <*mama, mamma*>, <*whisky, whiskey*>). A simple program removed such pairs, together with those such as <*fain, faun*> and <*groat, grot*> where both members are rare. The list also included some pairs of words which are almost synonymous, such as <*artist, artiste*>, <*babes, babies*>, <*waggle, wiggle*>. Although one of the members of such pairs might be considered more appropriate in a particular context, it does not seem to be a distinction that a computer spellchecker could be expected to make. As there was no way of identifying such pairs automatically, they were removed manually.

There were also some notable omissions from the list. Some of these were commonly confused pairs such as <*from, form*> (probably omitted because of the relatively high cost assigned to transpositions in Mitton's string-matching algorithm), and words containing apostrophes such as <*cant, can't*> and <*were, we're*> (apostrophes had not been considered by the list generation program). These omissions were rectified manually.

After several iterations of pruning and addition we rewrote the list with each pair appearing twice, both as an <*a, b*> pair and a <*b, a*> pair – both <*rite, write*> and <*write, rite*> were included in this list, for example. It was easier to use in this way since each word was in its alphabetical position as a word *a*. At this point there were around nine thousand pairs in the list.

1.1 Confusion functions

When scaling up the confusion-set approach, it becomes obvious that sets, strictly defined, are not really what you want. This is because of the great disparity in frequency between the members of some of the sets. Take the rare word *wold*, for example, which is in the same set as *world* and *would*. It makes sense to check every occurrence of *wold*, in case *world* or *would* was intended, but checking every *world* and *would* to see if they should be *wold* seems a waste of time and more likely to provoke errors than to correct them. What is required is an arrangement in which one word – here *wold* – is the headword, the one you want to check if you find it, and the others – *world* and *would* – are possible replacements for that word. That is to say, what is required is, strictly speaking, a function rather than a set. The classic confusion sets, of course, can easily be represented as confusion functions; if you want to check every *principle*, to see if it should be *principal*, and vice-versa, the set *{principle, principal}* simply becomes the two functions <*principle, principal*> and <*principal, principle*>. (The sets used by Mays et al. and the large confusion sets used by Fossati and Di Eugenio, referred to earlier, were also confusion functions in the sense just described, rather than classic confusion sets.)

The term "confusion set", however, is so well established in the literature that we will continue to use it, but the reader should bear in mind that, from now on, we are referring to confusion functions.

To create confusion sets from the pairs, each word *a* was taken as a headword and all the word *b*'s with which it was paired became its candidate replacements; the number of times each word appears as a word *a* represents the number of candidate replacements – so, for example, *write*, which is listed five times as a word *a* would end up with five candidate replacements.

To avoid the problem of widely disparate frequencies, pairs were first removed where the word *b* was rare

(occurring less than 80 times in the British National Corpus) and the word *a* common (occurring more than 8,000 times). Also removed were some pairs of confusables containing apostrophes, as these are commonly omitted but less often inserted (Mitton 1996). These included pairs where word *a* was a contracted form (e.g. *aren't, he'll, who're*). So, for example, the final list contained the pairs <*aunt, aren't*>, <*hell, he'll*> and <*whore, who're*> but not the pairs <*aren't, aunt*>, <*he'll, hell*> or <*who're, whore*>. Two exceptions to this were the commonly confused pairs <*its, it's*> and <*your, you're*> that were included both as <*a ,b*> and <*b, a*> pairs.

These two stages removed just over a thousand pairs. Creating sets from these pairs resulted in a total of nearly 6000 headwords with between one and five candidate replacements for each as shown in Table 1.

N. of replacements	N. of sets	Percentage
1	4461	75%
2	1063	18%
3	386	6.5%
4	30	0.5%
5	2	0.03%
Total Sets	5942	100%

Table 1: Confusion set sizes

The two with five candidate replacements were *sit* (*sat, set, shit, site, suit*) and *ware* (*war, wear, were, where, wire*).

The process described above may well be similar to the method used by Carlson et al. (2001) for creating their 265 confusion sets, which they describe as "using simple edit distance in both the character space and the phoneme space". They have kindly provided us with a version of their list. About a third of the confusion sets in their list are missing from ours. The great majority of these are inflected forms, such as <advance, advanced>. We decided not to include such sets in our list since, although (as we show later) it is a common error to write the base form of a noun or verb in place of an inflected form, it is trivial to generate such confusion sets on the fly, and, in the absence of a rationale for including some but not others, their inclusion would have enlarged the list many times over. Other missing ones were a small group of proper nouns and a few errors of grammar or usage, which were not of the type we were aiming to correct – <*among, between*>, <*fewer, less*>, for example.

Further detail is provided in (Pedler, 2007) and the final list of confusion sets is available from www.dcs.bbk.ac.uk/~jenny/resources.html.

2.A corpus of dyslexic real-word errors

We now had a list of confusion sets, but how could we assess whether our list contained the right sets, and enough of them, to make a worthwhile contribution to real-word error checking? Although artificial error data may suffice for experiments, it cannot answer questions such as these. We needed real-word errors, in context, produced by real people, in fact specifically dyslexics, since that was the primary focus of our research. Obtaining these proved to be difficult.

2.1 Compilation of a real-word error corpus

An earlier piece of work (Pedler, 2001) had collected about 600 dyslexic errors, of which 100 were real-word errors, but we clearly needed more. We contacted college disability officers, spoke with people who worked with dyslexics and posted to bulletin boards and mailing lists. Though many people expressed interest, only a few were able to supply actual examples, but we were finally able to assemble a corpus (the "base corpus") of over 21,000 words containing well over 2000 errors, of which 800 were real-word errors. Table 2 gives some summary statistics.

Sentences	1395
Words	21524
Total errors	2653
Real-word errors	833

Table 2: The base corpus

The most productive sources, contributing about half of the errors, were dyslexic bulletin boards and mailing lists on the internet. Other sources included essays and coursework by dyslexic students and some free writing by dyslexics collected by a research student for his PhD (Spooner, 1998).

Although the base corpus obviously contained non-word errors, they are just a distraction for a real-word error checker so a sub-corpus containing only the real-word errors was produced. Any sentences containing only non-word errors were removed and all non-word errors in the remaining sentences were replaced by their target words. (It is not unreasonable to suppose that, in actual use, a spellchecker could first apply a non-word error checker, since non-word errors are more straightforward to detect than real-word errors, and then make a second pass with a real-word error checker.) This sub-corpus (hereafter simply "the corpus") contained just over 12,000 words with a total of 833 real-word errors. It is available, with some documentation (see also Pedler, 2007), from www.dcs.bbk.ac.uk/~jenny/resources.html

It is likely that some of the text had already been spellchecked and, if so, that some of the real-word errors it contained had been generated by a spellchecker. Cupertinos are particularly likely to occur when a poor speller is presented with long suggestion lists. The intended word may not be in the list at all or, even when it is, it may be buried beneath a long list of obscure words. This is no help at all to a dyslexic, or anyone else, who didn't know how to spell the word in the first place. To stop the spellchecker complaining, they may simply resort to selecting the first word in the list. The varying proportion of non-word errors

in the different sources suggests that there had been some attempt at error correction in some of them. The essays were almost certainly spellchecked whereas no spellchecker was available to the participants in the research experiment and, although the bulletin board includes a spellchecking facility, users are possibly more concerned with getting their message across than with their spelling. Nonetheless, it is real-word errors that are our main focus here and, however they were generated, they formed part of real texts produced by real people trying to communicate.

The errors are marked up in the format illustrated below:

> The collation of the information was **<ERR targ = really>** relay **</ERR>** **<ERR targ = quite>** quit **</ERR>** easy to do.

This makes it a simple matter for a program to extract the errors and their corresponding target word from the corpus. It also enables an experimental spellchecker to ignore the target words when checking the text but at the same time to check the correctness of its suggestions.

2.2 Profile of the Real-word Error Corpus: Error Frequencies

Although the majority of these error words occurred just once as errors, a minority occurred repeatedly so that the number of distinct error types was approximately half the number of error tokens. The word occurring most frequently as an error was *there* with 40 instances of incorrect usage, followed by *to* with 27 instances. As a single error word can appear as an error for several different targets – for example, *quit* appears as a misspelling of both *quiet* and *quite* – the total number of distinct <error, target> pairs is higher than the total number of error types. This is summarised in Table 3.

Sentences	675
Words	12024
Total errors (tokens)	833
Distinct errors (types)	428
Distinct error/target pairs	495

Table 3: Composition of the real-word error corpus

Table 3 suggests that users have a tendency to produce certain misspellings consistently; Table 4 shows the frequency with which the error words occurred in the real-word error corpus.

Error words that occur in the corpus ten times or more are listed in Table 5. Many of these appear as an error for more than one target and so contribute to several of the distinct <error, target> pairs. Several of the short, high-frequency words in this list - *to, an, is,* - appear as errors for four or more different targets, which confirms earlier findings (Hotopf 1980, Sterling 1983, Mitton 1987) that a high proportion of real-word errors involve this type of word.

N. Occurrences	N. Error types
>10	10
6-10	10
4 or 5	17
3	25
2	48
1	318
Total error types	**428**

Table 4: Frequencies of error types in the corpus

Error	Frequency	N. targets
there	40	3
to	27	5
a	22	3
form	19	1
their	18	1
its	17	1
your	17	2
an	13	5
weather	12	1
were	11	2
cant	10	1
is	10	4

Table 5: Errors occurring 10 or more times in the corpus

Table 6 shows the frequency with which distinct error/target pairs occurred. Again, although most of the pairs occur just once, a minority occur repeatedly. In contrast to the findings for individual errors (Table 5) which showed that some words were often produced as a misspelling of several other words, this shows that there are some words which regularly appear as a misspelling of one other word in particular. Pairs such as these are likely to be good candidates for confusion sets; the ten most frequent are listed in Table 7.

Many of these top ten pairs also feature in the small list of sets of 'commonly confused' words used in much of the research discussed earlier. This confirms that, although the small number of these sets limits their usefulness for a comprehensive effort at real-word error correction, they do at least represent errors that users actually make.

Given large differences in word frequency, there is often a marked asymmetry in these common error-target pairs. Several of the pairs in Table 7 occur, in the corpus, only one way round, e.g. *college* is sometimes misspelt *collage* but never vice-versa. Table 8 lists the remainder, in which each word occurs both as an error and as a target, and shows that,

in all cases, one member of the pair appears as an error significantly more times than the other. The only pair that approaches interchangeability is *there* and *their*.

N. Occurrences	N. Pair types
>10	8
6-10	7
4 or 5	13
3	21
2	59
1	387
Total error pairs	**495**

Table 6: Frequency of error pairs in the corpus

Error|target pair	Frequency
there|their	35
form|from	20
to|too	19
their|there	19
a|an	18
its|it's	17
your|you're	15
weather|whether	12
cant|can't	10
collage|college	9

Table 7: Ten most frequent error|target pairs in corpus

Error|target pair	Count a|b	Count b|a
there|their	35	18
form|from	20	3
to|too	19	4
a|an	19	1
its|it's	17	5

Table 8: Those of the ten most frequent pairs that occur both ways round

2.3

2.4 Profile of the real-word error corpus: error characteristics

2.4.1 Homophones

Homophones are often used as the basis of confusion sets and feature prominently in the sets of commonly confused words used by many researchers. Six of the most frequent <error, target> pairs listed in Table 7 are homophones and, in total, 69 (14%) of the distinct error pairs in the corpus are homophones; those that appear more than twice are listed in Table 9.

Homophone set	N. Occs
there, their, they're	38
to, too, two	23
its, it's	17
your, you're	15
weather, whether	12
herd, heard	5
witch, which	4
hear, here	3
wile, while	3

Table 9: Homophone sets occurring more than twice in the corpus.

2.4.2 Simple errors

Mays et al. (1991) created confusion sets for a word by listing all the other words that differed from it by a single-letter insertion, omission, substitution or transposition. This method would generate the target for about two thirds of the real-word errors in this corpus; 63% of them differed from the correct word in just one of these ways (Table 10).

Error Type	N.Errors	Percentage Errors
Omission	142	29%
Substitution	104	21%
Insertion	56	11%
Transposition	12	2%
All simple	314	63%
All error pairs	495	100%

Table 10 : Proportions of simple error pairs in the corpus

2.4.3 Tagset types

A real-word error sometimes gives rise to a syntactic anomaly and this can be the basis of error detection. For this to be the case the error and the target must differ in their parts of speech. There are three possibilities to consider – words that have no part-of-speech tags in common (distinct tagsets), those where some but not all tags are in common

(overlapping tagsets), and those where all tags are the same (matching tagsets). Table 11 shows the number of pairs falling into each group. A syntax-based spellchecker could be expected to have reasonable performance with errors falling into the first group and perhaps have some impact on the second but obviously none on the third.

Tagsets	N.Errors	Percentage Errors
Distinct	327	66%
Overlapping	117	24%
Matching	51	10%
Total error pairs	495	100%

Table 11 : Count of error-target pairs by tagset type

2.4.4 Inflection errors

Many of the error-target pairs were noun-noun or verb-verb confusions, and further investigation found that a large number of these (20% of the total error pairs) were inflection errors. This corroborates a similar finding in Mitton's (1987) analysis of a corpus of school leavers' compositions (these were not the same as those included in this corpus).

About a third of the noun targets where the error was also a noun were number errors, almost exclusively a singular noun used in mistake for a plural. Many of these cases were simple omission errors resulting from the *-s* being left off the end of the word. Others (such as *virus* for *viruses* or *story* for *stories*) have a slightly more complicated plural form, but again the difference between the error and the intended word occurs right at the end. (The position of the error in the word is considered further below.)

Half of the verb targets where the error was also a verb involved a wrongly inflected form of the same verb. Many of these were regular inflections where the error involved the base form with an omitted *-s* (third person singular), *-ed* (past tense, past participle) or *-ing* (present participle). The remaining verb inflection errors were for irregular verbs, and, here too, the errors mostly involved producing the base form of the verb instead of the past tense or past participle.

2.4.5 Position of first letter error in word

A striking finding from Mitton's analysis of a corpus of errors from a university entrance exam (1996) was that real-word errors, including but not confined to inflection errors, tended to differ from their targets towards the end of the word – *though* for *thought*, *person* for *persons*, *notably* for *notable*, *word* for *world* – though this was not a feature of non-word errors in the same corpus. The same is true of this corpus; over half of the real-word errors differ from their targets at or near the end of the word. A secondary finding from Mitton's analysis was that a small group of real-word errors (again unlike non-word errors) differed from their target in the first letter, and this too shows up in this corpus, with 11% differing in the first letter, largely because of silent initials, such as *now* for *know*.

2.4.6 Proximity of errors

Real-word error checkers must use context in some way. Syntactic anomaly approaches to error detection generally use one or two words on each side of the suspect word to determine whether that word is improbable in that context. This will run into difficulty if these words are themselves errors. (Note again the inadequacy of artificial data, which tends not to contain this problem at all.) To assess the extent of this problem, we looked at the surrounding context for each real-word error. Table 12 shows the proportion of the real-word errors with another error (either a non-word or real-word error) within one or two words on each side. Admittedly, this corpus was taken from the writing of dyslexics, who make a lot more errors than most people, but it is people who have trouble with spelling who need a good spellchecker. For a quarter of the real-word errors in the corpus at least one other error occurs within two words to the left or right. In some cases (74 errors, 9% of the total errors) this is another real-word error. While a non-word error will be detected by dictionary look-up and the checker may be able to make an attempt at correcting it or at the very least will be aware that the context that it is considering is unreliable, another real-word error in the vicinity will compound the problem.

	Left	Right	Left & Right	Total
1 word each side	6%	7%	2%	15%
2 words each side	8%	12%	5%	25%
			errors = 100%	833

Table 12: Proportion of real-word errors with another error in the immediate context

3. How Would our List Cope with our Corpus?

We now have a large list of confusion sets and a corpus of real-word errors. Assuming we had some reasonably effective way of using this list to detect real-word errors, how successful might we hope to be? As described earlier (Table 3), the corpus contains 495 distinct <error, target> pairs, several of which occur frequently, giving a total of 833 errors overall. There are three possibilities:

- detectable and correctable: the error is the headword of a confusion set and the target is in its candidate list, so the spellchecker has some chance of detecting and correcting it;
- detectable but not correctable: the error is the headword of a confusion set but the target is not one of the candidate corrections, so the spellchecker might spot the error but it will not be able to suggest the correct replacement;
- not detectable: the error is not the headword of any of the confusion sets, so the spellchecker will simply ignore the error.

Inflection errors fall into the last two of these categories, because, as noted above, we had decided not to include pairs such as <*advance, advanced*> in our list of confusion sets. Some of these errors might be detected – if the error word

was the headword of a confusion set – but the great majority fall into the "not detectable" category. Although, in our opinion, they are probably not suitable candidates for the confusion-set approach, you could, of course, routinely check every singular noun and base-form verb, but this takes us away from the idea of using a predefined list.

Table 13 shows the proportion of the errors in the corpus falling into each of these categories both for error types (considering each error-target pair once) and for error tokens (the overall number of errors appearing in the corpus). A spellchecker using our list would have some chance of spotting 70% of the errors (tokens) but only for 58% could it also hope to supply the correction.

	Types	**Tokens**
Detectable and correctable	44%	58%
Detectable but not correctable	16%	12%
Not detectable (inflection error)	23%	17%
Not detectable (other)	17%	13%
Total (100%)	495	833

Table 13: Coverage of corpus errors

The majority of the pairs that the spellchecker would be unable to correct occur just once in the corpus. Those occurring more frequently are listed in Table 14. The striking thing about this list is the number of short function words it contains and the number of different permutations in which they occur, which again confirms earlier findings that these words are particularly problematic. The most frequently occurring of these pairs – <*a*, *an*> – should be easy to check for since the appropriate choice between them depends solely on whether the following word, when pronounced, starts with a vowel or a consonant. It is debatable, however, whether function words in general lend themselves to the confusion-set approach. *The*, the most frequent word in the language, appears in second place in the table with four occurrences as a misspelling of *they*. It also appears among the once-only pairs as a misspelling of *that* and *there*, suggesting that users have a tendency to produce *the* in place of other *th-* function words. But correct usages of *the* overwhelmingly outnumber the error usages and to check every occurrence as a potential error would be more likely to raise false alarms than to produce corrections.

It is questionable whether some of the words in the list should be considered as spelling errors at all – producing *i* for *it* is clearly a slip; *u* for *your* could be considered 'shorthand' of the type that is used in text messages; *cause* is probably intended as a colloquial version of *because*.

This leaves just three pairs that could be considered for future inclusion in the list of confusables – <*easy*, *easily*>, <*mouths*, *months*> and <*no*, *know*>. (*Mouths* for *months* may be a Cupertino, caused by people originally writing *mounths*.)

So, 70% of the errors are detectable in the sense that the error is the headword of one of our confusion sets; the spellchecker will consider them as potential errors, but what chance does it have of actually detecting them? To do this, it needs to apply rules based on the surrounding context. These rules can make use of any aspect of the text, syntactic or semantic, or any information about the language, such as word frequency.

Error not a headword ("non-detectable")		**Target not a candidate ("non-correctable")**	
Pair	**Frequency**	**Pair**	**Frequency**
a, an	17	an, a	4
the, they	4	cause, because	3
is, his	2	as, has	2
is, it	2	easy, easily	2
i, it	2	for, from	2
u, your	2	in, is	2
		mouths, months	2
		none, non	2
		no, know	2

Table 14: 'Non-detectable' and 'non-correctable' errors occurring more than once in the corpus

A syntax-based method is the most straightforward to implement and has been shown to have good performance in cases where the error causes a syntactic anomaly (Atwell and Elliott, 1987; Golding and Schabes, 1996). But of course it can only work if the error and the target differ in their parts-of-speech. Table 15 is restricted to the detectable errors, i.e. those where the error was the headword of a confusion set, and provides some encouragement for a syntax-based approach. For two-thirds of the errors, the error and the target do not share any tags in common, so a syntax-based approach should have a good chance of detecting the error; only for 7% would it have no chance at all.

Tagsets	**Types**	**Tokens**
Distinct	58%	68%
Overlapping	31%	25%
Matching	11%	7%
Total errors (=100%)	299	580

Table 15 : Comparison of tagsets of error and target, for detectable errors

These figures are, of course, upper limits on the detection of errors in this corpus with our list of confusion sets. Since we are not here evaluating any particular implementation of the confusion-set method, we are not saying how many of the errors would actually be detected. The same applies, and with greater force, to correction, i.e. the inclusion of the target in the list of suggested corrections

offered to the user. We can say that about four-fifths of the detectable errors could also be correctable, in the sense that the target is present in the candidate lists of those confusion sets, but we are not saying how often the target would actually be offered to the user, or how high up the list of suggestions it would appear.

It could be argued, however, that the question of correction is a secondary one for the confusion-set approach. Although the proposing of a correction has generally been regarded as an intrinsic part of the confusion-set approach and is, indeed, a natural by-product of this method – having decided that word *x* fits better than word *y*, it is natural to offer *x* as the correction – it is not necessary to use the same technique for both detection and correction. One could imagine a spellchecker, having decided through a confusion-set process, that a word is an error, simply passing the suspected error to a correction process similar to the one it uses for non-word errors. The spellchecker might use its goodness-of-fit results, gained from the confusion-set process, to enhance the correction part, but it need not be restricted to the candidates in the confusion set.

Conclusion

We have described the creation of a list of about 6,000 confusion sets suitable for a spellchecker employing the confusion-set approach to real-word error detection, and we have described a corpus of over 800 real-word errors culled from the writings of dyslexics. A spellchecker using this list of confusion sets might expect, as an upper limit, to detect 70% of the errors in the corpus. Many of the errors that would be undetectable with this list involve either function words or inflections. A spellchecker using a syntax-based approach should have a good chance of detecting the majority of the detectable errors.

The confusion-set approach, given a large list of confusion sets, can be expected to make a worthwhile contribution to real-word error checking, but it will fall short of being a complete solution.

References

Atwell,E. and Elliott, S. (1987). Dealing with Ill-formed English Text. In R. Garside, G. Leech and G. Sampson (Eds.), *The Computational Analysis of English*. London: Longman Publishers, pp. 120-38

Carlson, A.J. and Fette, I. (2007). Memory-based Context-sensitive Spelling Correction at Web Scale. In *Proceedings of the 6th International Conference on Machine Learning and Applications*, pp. 166-171.

Carlson, A.J., Rosen, J. and Roth, D. (2001). Scaling Up Context Sensitive Text Correction. In *Proceedings of the National Conference on Innovative Applications of Artificial Intelligence* pp. 45 – 50.

Fossati, D. and Di Eugenio, B. (2008). I saw TREE trees in the park: How to correct real-word spelling mistakes'. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pp. 896-901.

Golding, A.R. (1995). A Bayesian Hybrid Method for Context-sensitive Spelling Correction. In *Proceedings of the Third Workshop on Very Large Corpora*, pp. 39-53.

Golding, A.R. and Roth, D. (1999). A Winnow based approach to context-sensitive spelling correction. *Machine Learning* , 34 (1-3), pp. 107-30.

Golding, A.R. and Schabes, Y. (1996). Combining Trigram-based and Feature-based Methods for Context-sensitive Spelling Correction. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 71-78.

Hotopf, N. (1980). Slips of the Pen. In U. Frith (Ed.), *Cognitive Processes in Spelling*. Academic Press, pp. 365-404.

Jones, M.P. and Martin J.H. (1997). Contextual Spelling Correction using Latent Semantic Analysis. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 166-173.

Levenshtein, V.I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics – Doklady*, 10(8), pp.707-10.

Mays, E., Damerau, F.J. and Mercer, R.L. (1991). Context Based Spelling Correction. *Information Processing and Management*, 25 (5), pp. 517-22.

Mitton, R. (1987). Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers. *Information Processing and Management* 23(5) pp.495-505.

Mitton, R. (1996) *English Spelling and the Computer*, London: Longman Publishers.

Pedler, J. (2001) Computer Spellcheckers and Dyslexics - a Performance Study. *The British Journal of Educational Technology*, 32(1), pp. 23-38.

Pedler, J. (2007) *The Computer Correction of Real-Word Spelling Errors in Dyslexic Text*. PhD Thesis. University of London, Birkbeck.

Spooner, R.(1998). *A spelling aid for dyslexic writers*. PhD thesis, University of York

Sterling, C.M. (1983). Spelling Errors in Context. *British Journal of Psychology*, 74, pp. 353-64.

Wagner, R.A. and Fischer, M.J. (1974). The String to String Correction Problem. *Journal of the A.C.M.*, 21(1), pp. 168-73

Veronis, J. (1988). Computerized Correction of Phonographic Errors. *Computers and the Humanities*, 22, pp. 43-56