How Helpful Are Online English Learners' Dictionaries in Dealing with Misspellings?

Robert Lew: *Adam Mickiewicz University in Poznań, Poland (rlew@amu.edu.pl)*

Roger Mitton: *School of Computer Science and Information Systems, Birkbeck, University of London, UK (r.mitton@dcs.bbk.ac.uk)*

Abstract

This study looks at how well the leading monolingual English learners' dictionaries in their online versions cope with misspelled words as search terms.[1] Seven such dictionaries (*Longman Dictionary of Contemporary English*, free online version; *Longman Dictionary of Contemporary English*, premium subscription version; *Merriam-Webster's English Learner's Online Dictionary*; *Macmillan English Dictionary Online*; *Cambridge Advanced Learner's Dictionary*; *Oxford Advanced Learner's Dictionary*; and *Google English Dictionary*) are tested on a corpus of misspellings produced by Polish, Japanese, and Finnish learners of English. The performance of the dictionaries varies widely, but is in general disappointing. In a large proportion of cases, the dictionaries fail to supply the intended word, and when they do, they do not place it at the top of the list of suggested alternatives. Results are then compared with those from one year ago for the same dictionaries and the same misspellings. A detailed analysis follows, identifying some of the mechanisms behind the failures in identifying and correcting misspellings. The success rates of the dictionaries are compared with that of an experimental context-free spellchecker developed by the second author, and the spellchecker is found to be markedly superior. The data are subjected to a cluster analysis to see if the dictionaries can be grouped based solely on their performance. The article concludes with suggestions on how to improve the performance of the spellchecking facilities in online dictionaries.

1. The role of spelling in dictionary consultation

A painful limitation of traditional paper dictionaries — at least for the most popular type of semasiological dictionaries for languages with alphabetic writing systems — is that the primary access route requires the user to be familiar with the access alphabet of the dictionary (Nielsen 1995), and to know how the target item is spelled. With reference to the first point, users of modern electronic dictionaries are indeed (if only up to a point) 'liberated from the straitjacket of ... alphabetical order' (Atkins 1996: 516). However, the second point remains a valid concern: the user still needs to know how to spell the target word, or at least to enter something sufficiently close that the dictionary can find the required entry.

Of course, dictionary users cannot always be expected to replicate standard English spelling. Sometimes they make typos — 'performance errors'; sometimes they just don't know the correct spelling and they make a guess (or they think they know it but they're wrong) — 'competence errors'. For competence errors in particular, misspelling patterns typical of native speakers of English may be different from those of learners of English (Mitton and Okada 2007).

Further, online dictionaries are increasingly used in conjunction with online work and entertainment, such as when learners of English attempt to look up a word which they hear being

spoken while watching a TV show on their computer. Such a lookup situation is bound to generate queries where the search term, rather than representing a specific vocabulary item from the learner's lexical repertoire, is a 'creative spelling', a shot-in-the-dark, a transcription of what the user imagines he has heard. This is a little similar to what some call *phonetic spelling* (cf. Proctor 2002), but more complex, as here not one but at least two phonological systems are involved, each with its own phonotactic regularities and spelling-to-sound correspondences. We would expect the best electronic dictionaries to be able to offer useful assistance in all of the above cases, but do they actually provide such assistance?

2. Spelling correction in e-dictionaries

In a common type of e-dictionary interface, the user types a search term into a box; the dictionary has to find an entry corresponding to the search term or possibly (in more sophisticated dictionaries) an entry in which the search term occurs as part of a multi-word unit. An exact-match algorithm would assume that dictionary users are perfect spellers, which is obviously not a realistic assumption. A good dictionary interface should be able to guess the user's intention even if they misspell the word. However, in a recent analysis of three online German dictionaries (Bank 2010), only one dictionary has been found to be at all 'rechtschreibtolerant' — that is, able to deal with misspellings in any useful fashion.

A good dictionary interface, when presented with an unknown string, should make reasonable suggestions as to what the user may have meant. Furthermore, the guesses should be presented as an ordered list, with the best guesses at the top. Ideally, the one word actually intended by the user should be presented at the very top, but this ideal is not always achievable, even in the best possible system, due to the wide variation in misspellings.

First, the dictionary needs to recognize that the search term entered is not a standard spelling. Then, it needs to home in on a compact set of the most likely alternatives and rank them, so that they can be presented back to the user as an ordered list. Or, less commonly, it might just take the user to the entry for the top-ranking alternative (much as the Google search engine currently does). The procedure is similar to that involved in spellchecking, but there are differences: the online dictionary lacks the benefit of context but, on the other hand, it does not need to handle punctuation, numerals, obscure proper nouns and so on.

2.1. Types of spelling errors

Many of the spelling errors in running text are single-letter departures from the target word. Taking the target word *trepidation* as an example, these fall into one of the following four subcategories: a single letter is omitted (*tepidation*); a single letter is wrong (*trepitation*); one extra letter is inserted (*treppidation*); two adjacent letters are transposed (*trepidaiton*). According to some studies (Damerau 1964; Pollock and Zamora 1984), such simple errors may account for over eighty percent of misspellings. However, this percentage is likely to be lower with a more realistic representation of poor spellers in the corpus: sixty-nine percent in Mitton (1996: 46). Many (though not all) of these simple errors tend to be the result of mistyping words. As such, they are mechanical errors of performance, rather than errors of competence, and some authors use the term *misspelling* in a narrower sense which excludes mistypings (e.g. Deorowicz and Ciura 2005). Though it is not always possible to categorize an error as one type or the other (e.g. **accomodation* for *accommodation*, or

**consistant* for *consistent*), their underlying causes are different. It is misspellings of the competence type that are our primary focus here.

At the other end of the mechanical-conceptual cline, there are non-standard formations at the lexical-morphological level, such as when a speaker actually has the word **unpolite* in their mental lexicon and uses it in place of (or as a variant of) the standard *impolite*. Though sometimes the source of genuine problems, especially for non-native users of a language, it is doubtful if such errors of lexical competence should be classified as strictly *spelling-related* (pace Deorowicz and Ciura 2005).

2.2. Rare versus common words

For a spellchecker processing running text, it is reasonable to assume that an instance of a rare word (especially a *very* rare one) may be a misspelling, particularly if there is a common word to which it bears some similarity. For example, as pointed out by Mitton (1996: 96), the orthographic string *wether* when found in running text is more likely to be a misspelling of either *whether* or *weather* than the rare word meaning 'a castrated ram'. Spelling correctors working with text can use this information to detect and flag such potential real-word errors. However, in a corpus of strings being looked up in an online dictionary, the frequency distribution of word forms is less skewed than in running text (De Schryver et al. 2006), so that even quite rare words have a fair chance of being looked up. This makes perfect sense: when someone reads a text, they will not usually be troubled by the common words, but the occasional rare word is likely to be looked up. So, although De Schryver et al.'s study of log files presents only a single piece of evidence, it is reasonable to assume that native speakers, and to a lesser extent advanced learners, often consult their dictionaries for less frequent words.

2.3. The role of context

A spellchecker checks running text, and some of the more advanced systems attempt to utilize the context, both to detect the misspelling – most real-word errors can only be detected by using context – and to refine the list of suggestions. However, when someone types a word into the search window of an online dictionary, no contextual information is available to the dictionary application. Still, most spellcheckers designed for the correction of texts do not use context either, and yet achieve good success rates nevertheless (Kukich 1992; Deorowicz and Ciura 2005; Mitton 2009).

3. The study

3.1. Aim

The aim of the study is to assess the performance of the leading monolingual learners' dictionaries of English in their online versions at guessing the required word when presented with misspelled versions produced by foreign learners. Often there will be several plausible alternatives, so dictionaries will customarily provide not just one suggestion but a short list. In such a case, the nearer the top of the list the intended word appears, the better the performance of the spelling correction mechanism.

More specifically, we would like to find out whether the level of performance of the most prestigious dictionaries is in general satisfactory, to what extent the different dictionaries perform similarly or differently, and how specific dictionaries compare with the others.

In view of the preliminary results indicating that the tested dictionaries performed below expectation, a further aim was added during the course of the study, and for this, the original author was joined by the second author. This further aim was to see if an experimental context-free spelling corrector designed by the second author (Mitton 1996) would be able to perform better than the dictionaries tested.

3.2. Corpora of misspellings used in the study

The corpus of spelling errors used in the present study consists of 200 attempts at spelling English words by native speakers of languages from three different language families: Polish (100 items), Japanese (50), and Finnish (50). A brief description of the three sets of misspellings follows, and a sample of ten items from each is given in the Appendix.

3.2.1. Polish misspellings. The largest part of the corpus was made up of misspellings by Polish writers, collected in 2010 by the first author, with the help of two student assistants as experimenters.

The data were collected by way of oral elicitation. A set of English words known to be frequently misspelled was taken from *The 200 Most Commonly Misspelled Words in English*[2] reported by Richard Nordquist. One by one, the words from the list were played back in audio form to one of two Polish learners of English in their first year of college (one female from Szczecin University, one male from Gdańsk University), using the built-in audio pronunciation capability of the popular bilingual English-Polish dictionary Diki.pl, known for its decent audio quality. Thus, a target word would be played back to the participant without disclosing its spelling, and the participant would respond by typing the word into the computer. The experimenter would wait until the participant indicated that they were done, and then proceed to play back the next target word. Participants had been instructed in the warm-up sessions to proceed as if they were using an online dictionary to look up words they had just heard.

All the typed word-like strings were logged. Correctly spelled words as well as obvious mistypings, which in all likelihood would not have challenged the spellchecking algorithms of the dictionaries, were subsequently removed, with the remaining strings yielding the Polish subcorpus of 100 misspellings.

3.2.2. Japanese misspellings. The 50 Japanese misspellings were taken from the SAMANTHA Error Corpus created by Takeshi Okada at Tohoku University, Japan (Okada 2005). Japanese university students were asked to write down a series of English words. For each one they were given its definition in Japanese and an approximate representation of the English pronunciation in the Japanese moraic (or, more loosely, syllabic) script katakana. For the present study, we confined our attention to those of the Japanese misspellings that contained more than one single-letter error (and thus would provide a greater challenge for spellcheckers), selecting, for each target word, the most common of these. Up to a point, though perhaps not as much as for the Polish sample, the elicitation technique used would be likely to produce misspellings influenced by the typical sequencing of letters and sounds in Japanese, as well as by spelling-to-sound correspondences in English.

3.2.3. Finnish misspellings. The Finnish data were collected by Suomi (1984) as part of her MA research. The errors were taken from test papers written by 60 Finnish speakers, aged 15-16 years, who had had about 16 hours per week of English at school for six or seven years. There were two tests. In the first, the students were presented with a short written dialogue, mostly in English but with

some sentences in Finnish; they had to write their translations of these sentences. In the second, the students listened to a short dialogue, in English, then wrote their answers, in English, to questions (in Finnish) about the dialogue.

The set of Finnish misspellings is one of several included in the Birkbeck spelling error corpus (Mitton 1985) available from the Oxford Text Archive. (The data collected by Suomi also includes misspellings from native speakers of Swedish, but, for the present study, only the data from native speakers of Finnish were used.) Trivial errors were discarded, as they were in the case of the Polish subcorpus. This resulted in a list of 50 misspellings.

3.3. Dictionaries tested

Each of the misspelled words in the corpus was looked up in each of the following seven online dictionaries, all except the Google Dictionary being dictionaries for advanced learners of English, and all but one available at no charge. The seven dictionaries tested were (their URL's are given in the **References** section):

1. *Longman Dictionary of Contemporary English*, free online version (henceforth, LDOCE Free);
2. *Longman Dictionary of Contemporary English*, premium subscription version (LDOCE Premium);
3. *Merriam-Webster's English Learner's Online Dictionary* (MWALED);
4. *Macmillan English Dictionary Online* (MEDO);
5. *Cambridge Advanced Learner's Dictionary* (CALD);
6. *Oxford Advanced Learner's Dictionary* (ALD); and
7. *Google English Dictionary* (GoogleED).

The general idea was to test English monolingual dictionaries for learners of English available freely online. The set of leading English monolingual learners' dictionaries is actually well defined, and is frequently referred in the lexicographic literature as the Big Five, and includes: ALD, LDOCE, COBUILD, CALD, and MEDO. Of these, COBUILD has not been tested as it does not currently offer a free online version. For LDOCE, two versions were tested: the free online version, and also a Premium version. This version is available by subscription, with time-limited access granted to buyers of paper and DVD-Rom copies. It was included in order to see if paying users were being served better than users of the free version. (In fact, quite the reverse turned out to be the case, as we shall see below.)

In addition to these four British learners' dictionaries, we also included MWALED. Even though in terms of lexicographic content this American-made learner's dictionary may still not compare very favourably with the Big Five (Hanks 2009; Bogaards 2010), its web interface does offer some commendable features (Lew 2011).

Finally, GoogleED was also included in the study. GoogleED used to be a learners' dictionary of sorts, with the core lexicographic content apparently based on COBUILD. In August 2010, GoogleED switched over to the *Oxford American College Dictionary* (Lindberg 2006), which is not a dictionary targeted at language learners, but primarily at American college students speaking English as their native tongue. However, four factors spoke in favour of including GoogleED in the sample.

First, being associated with Google, the unquestioned leader in search engines, it was reasonable to expect it to become a very significant player also as an online dictionary of English for non-English-speaking users.

Second, its history as an online version of COBUILD, one of the Big Five, is in itself significant, and may have attracted a number of learner users who remained regular users even after the switch.

Third, although the *Oxford American College Dictionary* is a native-speaker dictionary, it is largely based on the *New Oxford American Dictionary* (McKean 2005), which, in turn, grew out of the *New Oxford Dictionary of English* (Hanks and Pearsall 1998). This latter dictionary benefited from Patrick Hanks' prominent involvement with the COBUILD project, and so in many ways is closer to the learner dictionary model than a traditional dictionary for native speakers of English.

Finally, given Google's prominence as a virtual synonym for data search and access, we were interested to see if GoogleED would perform better than the 'regular' dictionaries.

In August 2011, Google discontinued the autonomous GoogleED interface without as much as a word of warning or explanation. However, as of February 2012, the GoogleED can still be accessed by using the *define: term* syntax in a general Google search, and then clicking on *more* within the top item on the results list, which selects the *Dictionary* tab from the sidebar on the left of the Google search user interface. Admittedly, this is a lot of clicking that will discourage all but the most determined users, but the same effect can be achieved more directly by appending a parameter value of *tbs=dfn:1* to a Google search.[3] Despite its going underground, as it were, we have decided to recollect data from GoogleED for this study in an effort to have as much comparability as possible with Lew and Mitton (2011).

3.4. Procedure

All lookups were performed manually online by the first author, between January 18 and 21, 2012. For each misspelled word, the misspelling was pasted into the search box of each of the dictionaries. In every case, it was noted whether the dictionary was able to identify the correct target word, and, if the dictionary provided a list of alternatives, what was the position of the target word relative to other suggestions. The word (or non-word string, as was sometimes the case) presented at the top of the suggestions list was also noted, as well as any other striking suggestions further down the list.

As an illustration of the procedure, consider **Figure 1** below, taken from a test lookup in CALD. The intended word was *temporary*, and it was misspelled as **tempori*. The dictionary returned a list of ten suggestions. The top suggestion (number 1 on the list) was *temporise*, which was not the intended word. However, the correct target word *temporary* was found further down the list: in this case it was listed ninth. So, position 9 was noted for this misspelling in CALD.

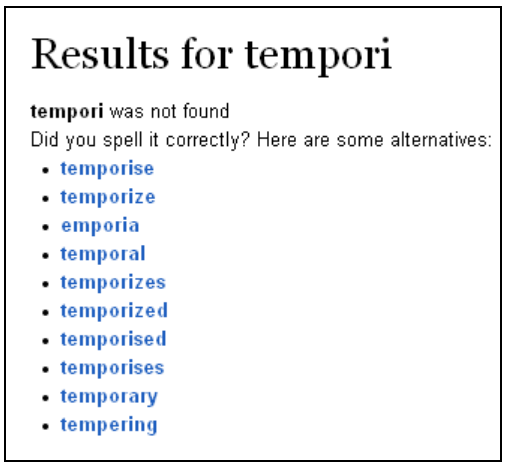

Figure 1: Example suggestions list in CALD for the target word *temporary* misspelled as **tempori*.

This example is representative of six of the seven dictionaries tested; the one exception was GoogleED, which did not provide a longer list of suggestions, but only a single alternative (if any at all). For some items for which it could not match an entry in the proper dictionary, GoogleED gave alternative 'Web definitions'. These alternative suggestions came from external online glossaries and encyclopedias (including Wikipedia) and were ignored in our evaluation.

Data for all dictionaries and misspellings were keyed into a database and analyzed so as to evaluate the relative performance of the seven dictionaries.

3.5. How well the dictionaries performed

Results for the complete corpus of 200 misspellings are presented in **Table 1** and **Figure 2** below. Percentage figures in the table cells indicate what proportion of the 200 target words were found in the respective positions within the individual suggestions lists returned by the dictionaries.

The figures under the heading *First* cover those cases where the target word was presented at the very top of the list of suggestions. *Top 3* means that the target was listed as first, second, or third, and so on. These figures are cumulative, so if the target was listed at the top of the list, it was automatically counted under all four categories (i.e. *First*, *Top 3*, *Top 6*, and *Top 10*). **Figure 2** conveys the same results in graphic form.

Table 1: Success rates for the seven dictionaries across all data. Figures indicate the proportion of target words found in the respective positions in the suggestions list.

	Target word listed in position:			
Dictionary	First	Top 3	Top 6	Top 10
LDOCE Free	51%	64%	74%	77%
MWALED	47%	57%	63%	66%
LDOCE Premium	50%	59%	60%	62%
CALD	35%	50%	55%	57%
MEDO	24%	43%	51%	54%
ALD	25%	43%	47%	51%
GoogleED	43%	(43%)	(43%)	(43%)

Two things are immediately obvious in the results: the wide variation between the different dictionaries, and the disappointing performance of most of the dictionaries tested. It is worth remembering here that our corpus of misspellings was designed to be challenging. Unlike some other studies, we did not focus on typos, most of which are simple errors that can be corrected with unsophisticated algorithms. Still, the very wide disparities between the success rates do indicate that at least some dictionaries are not doing the best job possible, to put it mildly.

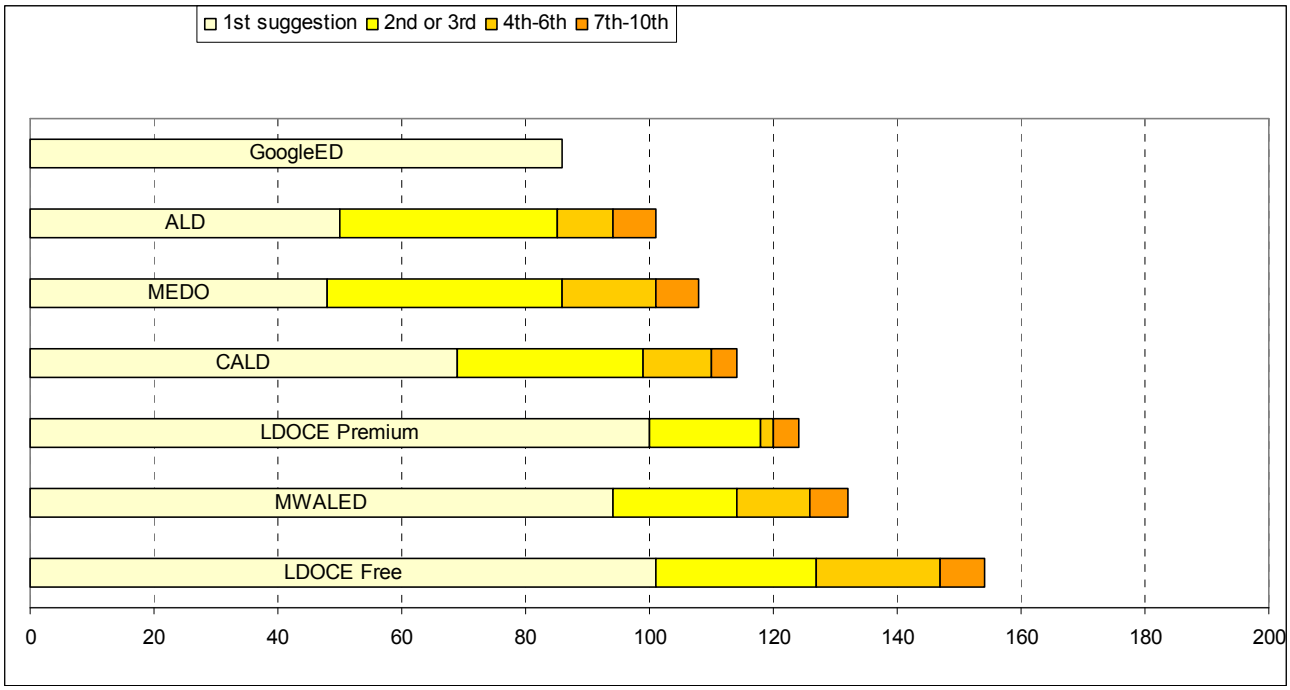

Figure 2: Performance of the seven dictionaries for all data (N=200). Bar sections indicate the number of target words ranked in the respective positions in the suggestions list.

Looking at the success rates for the first suggestion (top of the list), there is a gap between the two versions of LDOCE and MWALED on the one hand and ALD and MEDO on the other, with CALD in between. LDOCE and MWALED succeed in placing the target word at the very top of the list of suggestions about half of the time, with LDOCE being marginally better than MWALED; GoogleED does only slightly worse than LDOCE and MWALED in this respect. In contrast, ALD and MEDO only get about a quarter of the target words at the top of the list, and CALD about a third.

If we now lower the bar and include all suggestions in the top ten, then ALD and MEDO catch up somewhat, largely thanks to being able to include more of the target words in second or third place. But even with the top ten items on the list included, these dictionaries only succeed in 51% and 54% of the cases, respectively. On the top-ten measure, MWALED (66%) gets slightly ahead of LDOCE Premium (62%), but it is LDOCE Free that really surges ahead (77%), with a lot of accurate guesses in its lists found between the ranks of 2 and 6. It clearly outperforms all the other dictionaries, including, surprisingly, its deluxe sister LDOCE Premium.

We did not find any interesting differences in performance depending on whether the misspellings came from the Polish, Finnish, or Japanese subcorpus. The interested reader is referred to Lew and Mitton (2011) for some details.

3.6. Have the dictionaries improved, compared to one year before?

Since we already had data available for the same seven dictionaries and the same corpus of misspellings collected exactly a year earlier (between January 16 and 19, 2011), we thought it might be interesting to compare the performance of the dictionaries at the two dates. The differences are shown in **Figure 3**. For each dictionary, we give three measures, based on the position of the target word in the list of suggestions. The bars on the left represent items showing improvement: in these cases the target is now placed higher in the list of suggestions than it was in January 2011. The middle

bars stand for items with no change. The bars on the right indicate items where the target word has fallen down in the suggestions list.

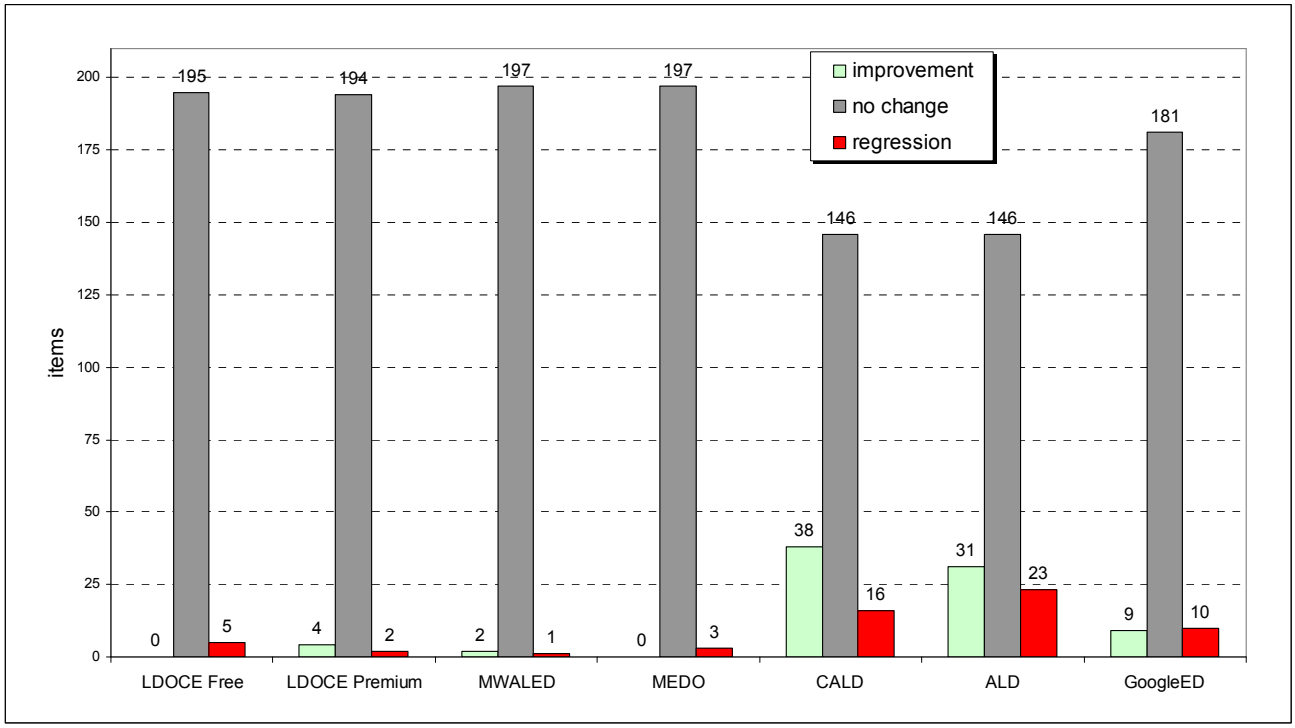

Figure 3: How the seven dictionaries changed between January 2011 and January 2012. The three bars for each dictionary represent, from left to right: items for which results have improved; items with no change; and items which got worse.

Of the seven dictionaries, there has been relatively little change for LDOCE Free, LDOCE Premium, MWALED, MEDO, and some change for CALD, ALD and GoogleED. Only two dictionaries have improved markedly: ALD and (especially) CALD. Whatever changes were implemented in the intervening year, they resulted in a greater number of target items guessed more accurately, though it is disappointing that, in both dictionaries, a non-trivial proportion of items have regressed since January 2011. In LDOCE Free and MEDO very few target words changed their standing on the suggestions lists, but unfortunately all those isolated instances were changes for the worse. In LDOCE Premium and MWALED there were very few changes either way, with a very marginal advantage on the side of positive changes. In GoogleED there were quite a few items with changed results, but the net effect is marginally negative.

This paints a rather gloomy picture. Most dictionaries have not improved over the year, and there have been quite a few setbacks. Of course, summary measures do not tell the whole story, and we will now proceed to the details.

3.7. Where the dictionaries failed

Since we have complete records of the suggestions offered by the respective dictionaries, in this section we will try to identify particular shortcomings in their performance. We offer some comments as to what may have caused the less-than-optimal guesses, and how these could have been avoided, indicating, where relevant, what has changed between January 2011 and January 2012.

Starting with the ALD, it seems this dictionary attaches too much weight to substring matching. This might explain why it would offer *apology* for **sakology* (a phonetically-motivated misspelling of *psychology*). Apparently, the dictionary homes in on the *-ology*, and then repeats the process with what remains, finding that *ap-* and *sak-* share the letter 'a'. The remaining items on the suggestions list are as follows: *sexology*, *sinology*, *biology*, *geology*, *ufology*, *enology*, *zoology*, *horology*, and *tautology*, in this order. Compared to 2011 results, a few items on the lists have switched places, and a couple of suggestions have been replaced with more obscure ones (e.g. *ecology* is out, *enology* is in), but the general problem persists. In the 2011 evaluation, ALD did not seem to give much regard to the first letter, even though research has shown that people generally get the first letter right (Yannakoudakis and Fawthrop 1983; Mitton 1996). For instance, the dictionary used to offer *deferens* for **referens* (*reference*). In the new evaluation, this aspect has improved, and ALD now correctly guesses *reference*.

A particular oddity of the suggestions served up by MEDO, and occasionally also ALD and CALD is their tendency to offer words with an *–s* at the end, even though there is no indication in the misspelling that one is required. Thus, all three dictionaries put *recommends* as first choice for **rekomend*, with the correct *recommend* only appearing in second place. Similarly, for the easy misspelling **disapoint*, we get *disappoints* at the top of the list. In 2011 all three offered *citizens* for **sitizen*, *forwards* for **fowards*, *repetitions* for **repetyszyn* and even *spaghettis* for **spagetti.* Compared with 2011, ALD and especially CALD have improved for some of such cases, so now, of the three, only MEDO favours *citizens,* ALD and CALD no longer offer *forwards*, nor does CALD offer *repetitions*, and none of them offers *spaghettis*. The changes compared to January 2011 are going in the right direction, but nevertheless the question remains why the other such cases still persist (and the plurals still often appear in second place, which they probably do not deserve). This mysterious tendency loses the three dictionaries quite a few easy points for top suggestion, at the same time inflating their *top 3* counts, as the reasonable suggestion tends to appear second in such cases. Why would all of ALD, CALD, and MEDO be affected by this overeagerness to tag on *–s*? Perhaps this has something to do with the software for dictionary compilation and publication that all three use IDM PitchLeads as the online platform (Glennon, personal communication). However, as far as we know, LDOCE also uses the IDM system, and yet it does not exhibit the *–s* problem. (We shall explore empirically the degree of relatedness between the dictionaries in Section 5.)

Conversely, MEDO, ALD and CALD all place the singular *university* at the top of the list of suggestions for **univercitys* (a misspelling of *universities*). For once, the plural would have been reasonable, and yet it is only given in second place.

At times, the suggestions offered by our dictionaries can be downright bewildering. A case in point are MWALED's offerings for **das*, a misspelling of *does*. Admittedly this is a challenging item, but the suggestions are puzzling, to say the least. The dictionary's output is given in **Figure 4** below, and it comprises three suggestions: *cream soda*, *giant panda*, and *piña colada*. Only a closer look at the entry can reveal why MWALED should come up with such a list. The plural for these compounds is given in a traditional compressed form as '~-das', and apparently it is this string that the dictionary has homed in on. Another surprise from MWALED, though this time with no apparent explanation, is the suggestion *archdiocese* for **ridicyles* (a misspelling of *ridiculous*).

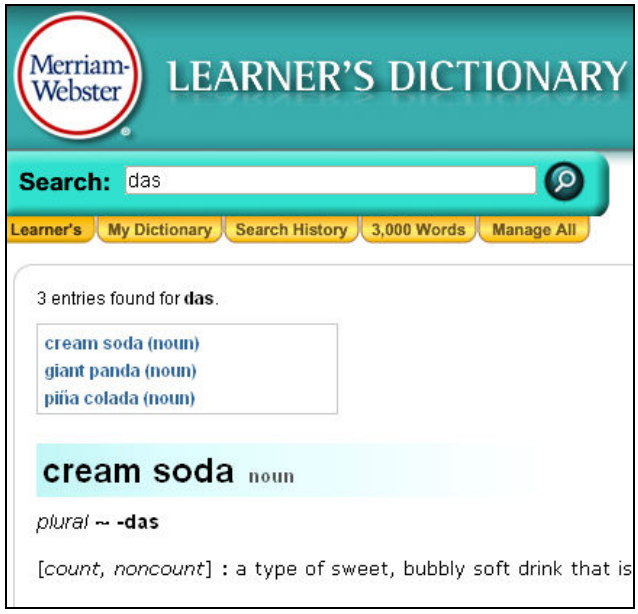

Figure 4: MWALED's suggestions for **das*, a misspelling of *does*.

A further mystery about MWALED's performance is the problem it has with the misspelling **spagetti* — possibly the easiest item in the whole corpus, which all the others get right. MWALED offers here no less than 16 alternatives (*spigot*, *spectate*, *spotted*, *spotlight*, *speculate*, *spectacle*, *septet*, *aseptic*, *sabotage*, *septic*, *sceptical*, *sceptic*, *seepage*, *sceptically*, *slippage*, *spatula*), but the obvious *spaghetti* is not among them, even though, to be sure, the entry for it is in the dictionary.

MWALED's algorithm seems to focus excessively on transpositions — it tends to rearrange the original letters: it offers *heir* for **hier* (*here*), *grade* for **gread* (great) and *crane* for **crean* (*clean*). All of the above problems, first identified in 2011, persist into 2012.

Life is made difficult for a spellchecker if its dictionary contains peculiar entries. This is to some extent true of all our dictionaries, but especially of ALD and GoogleED. In the absence of any data on word frequency — and they do not seem to be using any — these odd words just enlarge the set of (apparently) plausible corrections, and so we find the following unhelpful suggestions among the 'best' guesses in ALD: *etyma*, *xylem*, *inf*, *umbrae* (two more have been fixed since 2011), as well as proper names like *Tok Pisin* and *Wat Tyler* (one fixed). On top of that, GoogleED would not infrequently provide non-words among its suggestions, often only partially closer than the misspelling to any real English words. Even though four of such bizarre suggestions have disappeared since January 2011, the following still remain: **sejfy* for **sejfty* (*safety*), **sinirli* for **sinsirli* (*sincerely*), **bicikli* for **beisikli* (*basically*), and **identiti* for **aidentiti* (*identity*).

The *–ing* ending seemed to be another cause of difficulty for these dictionaries. In January 2011, only GoogleED was able to correct **useing* to the intended *using*. In January 2012, LDOCE Premium and ALD also get it right. But LDOCE Free and MEDO offer *unseeing* (a rare word but not entirely implausible), MWALED *seeing*, and — strangest of all — CALD proposes the nonce form *useding* (see **Figure 5**), apparently as a hypothetical inflected form of *used to*, as this is the entry to which the user is taken upon clicking on *useding*.

Figure 5: CALD suggestions for the target word *using* misspelled as **useing*.

Another easy case is **diging*, a straightforward misspelling of *digging*. As for *useing* above, GoogleED gets it right, and so does ALD, and now also CALD (in January 2011 it had the curious *ziging*). LDOCE Free still suggests *dining* (and, in third position, *diggings*, but never *digging*). LDOCE premium would rather have *dodging*, MWALED insists on *Diegan*, and MEDO would like *dinging*.

A rather striking feature of LDOCE (especially the free version) is that it likes to make two correct words by sticking a space in the middle of the misspelling, thus: *of fen* for **offen* (*often*), *inter fir* for **interfir* (*interfere*), *so rid* for **sorid* (*solid*), *back en* for **backen* (*bacon*), *be course* for **becourse* (*because*), *ail and* for **ailand* (*island*). This strategy may be occasionally successful when checking running text, but it does not work well for isolated dictionary query strings, especially if the spellchecker does not care whether the resulting pair is a likely combination.

Apart from that, LDOCE's offerings, among the dictionaries tested, tend to be the most respectful of the misspellings, with the suggestions generally retaining the first letter and the general word structure.

A new development, not noted in January 2011, is the tendency of ALD and CALD to suggest compounds spelled as separate words and phrasal verbs rather than the reasonable simplex forms. For example, both dictionaries place *dining car* at the top of the list for **dyning* rather than the simple and correct *dining*. **vater* (for *water*) gets *later on* (CALD) and *cater to* (ALD) as the best suggestion, and both dictionaries follow down the list with an assortment of compounds with water (*water gun*, *water ice*, etc.). For **szajning* (a misspelling of *shining*), CALD gives *signing up* and then, third on the list, *training bra*. While it is true that multi-word items have in many ways been the lexicographic underdog, prioritizing them in cases like the above seems more than a little far-fetched.

4. Can the dictionaries do better? Mitton's experimental spellchecker

As the online dictionaries clearly performed below expectation, the first author wondered if there were context-free spellcheckers capable of doing better. A literature search identified a promising context-free experimental spelling correction system (Mitton 1996, 2009). The second author was contacted and he offered to run the same data through his spellchecker. (This was the version of the

spellchecker designed for native speakers of English, i.e. without any adaptations for speakers of other languages.)

4.1. How Mitton's spellchecker works

When presented with a misspelling, Mitton's spellchecker begins by assembling a collection of dictionary words — typically a few hundred — that somewhat resemble the misspelling. It then takes each of these candidates and matches it against the misspelling to assess how good a candidate it is. The string-matching algorithm is a version of the well-known 'edit-distance' algorithm (Levenshtein 1966; Wagner and Fischer 1974; Véronis 1988).

The algorithm calculates the minimum number of editing operations required to get from the candidate to the misspelling, where each editing operation consists of inserting a letter, or deleting a letter, or changing one letter to another. For example, if the misspelling was **sakology* and the candidate was *psychology*, you could get from the candidate to the misspelling by deleting the *p*, changing the *y* to an *a* and the *c* to a *k,* and deleting the *h* – a total of four operations, therefore an edit-distance of four.

Merely counting the edit operations, however, only takes you so far. Consider the candidate *ecology*. You can get from *ecology* to **sakology* by inserting an *s*, then changing the *e* to an *a* and the *c* to a *k* – an edit-distance of three. So, simply on the basis of the number of operations, *ecology* would be preferred to *psychology*. But this does not seem quite right.

The *p* of *psychology* is silent so it is not surprising that people sometimes omit it; the *y* is relatively unstressed, and people often make mistakes over unstressed vowels, and the *ch*, in this word, corresponds to the same sound as a *k*. By contrast, if you were trying to write *ecology*, starting with an *s* would be an odd thing to do.

We can accommodate this by assigning a cost to each editing operation, with more serious (i.e. less likely) operations having a higher cost. We might decide that the operations on *psychology* are relatively insignificant and assign a cost of just one to each of them. For *ecology*, we might, similarly, assign a cost of one to changing *e* to *a* and *c* to *k*, but a much higher cost, perhaps four, to the unlikely error of inserting an initial *s*. If we now adapt the algorithm so that it calculates the cost, rather than the number, of the editing operations, we come out with a cost of four for *psychology* and six for *ecology*, so we would present *psychology* higher up the list of suggestions.

The dictionary inside Mitton's spellchecker is primed with information about appropriate costs to use in the string-matching. These are based partly on pronunciation and partly on analyses of large corpora of misspellings. So the spellchecker already knows, so to speak, that you might omit the *t* of *mortgage* or the middle syllable of *remember*, that you might insert an *s* into *latest* (**lastest*), that you might begin *phantom* with an *f*, and so on.

Readers wanting more than this brief sketch are invited to consult Mitton (2009) or, for more detail, Mitton (1996).

4.2. Mitton's spellchecker versus online dictionaries

Table 2 compares the success rates (in the same fashion as in **Table 1**) of Mitton's experimental spellchecker with the best-performing online dictionary (LDOCE Free), and **Figure 6** compares it with all the dictionaries graphically.

Table 2: Success rates of the best-performing dictionary compared with Mitton's experimental spellchecker, for all data.

	Target word listed in position:			
Dictionary	First	Top 3	Top 6	Top 10
Mitton	73%	87%	91%	93%
LDOCE Free	51%	64%	74%	77%

Mitton's spellchecker was able to place the intended target word among the top ten of its list of suggestions for 93% of the misspellings. The best dictionary in our set, LDOCE Free, performed significantly worse, achieving a success rate of 77%. The gap is even greater if we consider the spellchecker's ability to place the target word in the most valuable top portion of the list of suggestions. Here the experimental spellchecker outperforms LDOCE Free by over 20 percentage points (both for *First* and for *Top 3*).

In comparison with the other dictionaries, of course, the gains are still greater (**Figure 6**). From another perspective, the experimental spellchecker was able to guess perfectly 23 items (by placing them at the top of the list of suggestions) which *none* of the seven dictionaries managed to get right.

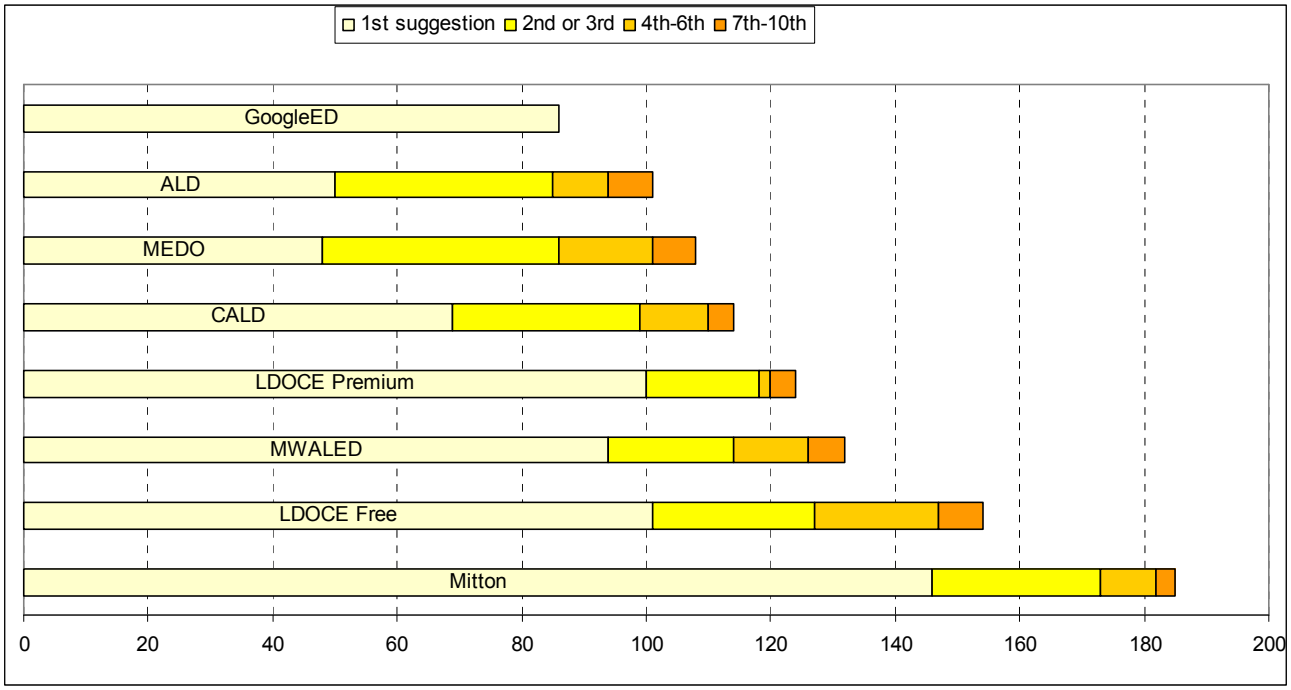

Figure 6: Performance of the seven dictionaries compared with Mitton's experimental spellchecker, for all data (N=200).

5. Similarities between the dictionaries

In section 3.7. above we saw over and over again dictionaries returning similar results, or indeed falling into the same traps. We hypothesized that some of the similarities may be due to the use of the

same online dictionary platform, PitchLeads from IDM, which, to our knowledge, is used by the two LDOCE dictionaries, ALD, CALD, and MEDO. This would leave only two dictionaries in our sample, MWALED and GoogleED, *not* using it. But there are other factors which could determine the ranking of words, such as the wordlist or particular techniques of spellchecking. In this section we would like to explore the similarities between the dictionaries in a more formalized way.

For the purposes of this study, the main parameter of interest is the position of the intended word on a list of suggestions. Thus, if two dictionaries both present the intended word at the top of the list, or if both list the word in the same position (say, third), this means the two dictionaries perform identically. Conversely, the greater the disparity between the ranks of the target word in two suggestions lists, the farther apart the dictionaries are. In order to quantify this measure, we computed pairwise Spearman's rank-order correlation coefficients for the dictionaries. Mitton's spellchecker was included , but GoogleED was not, because this dictionary only offered at best a single suggestion rather than a list, imposing a radical restriction on the range of possible values in an analysis of ranks. The figures are provided in **Table 3**. (The table is symmetrical about the diagonal since the correlation of A with B is, of course, the same as the correlation of B with A.)

Table 3: Pairwise Spearman correlation coefficients for target word rank data (N=200).

	Mitton	LDOCE Free	LDOCE Premium	MWALED	MEDO	CALD	ALD
Mitton	1.00	0.44	0.48	0.39	0.38	0.29	0.32
LDOCE Free	0.44	1.00	0.69	0.49	0.51	0.51	0.41
LDOCE Premium	0.48	0.69	1.00	0.46	0.54	0.50	0.44
MWALED	0.39	0.49	0.46	1.00	0.41	0.37	0.32
MEDO	0.38	0.51	0.54	0.41	1.00	0.75	0.74
CALD	0.29	0.51	0.50	0.37	0.75	1.00	0.75
ALD	0.32	0.41	0.44	0.32	0.74	0.75	1.00

It is evident from the correlation coefficients that some dictionaries indeed exhibit greater affinity than others. LDOCE Free, for instance, correlates most highly with LDOCE Premium at 0.69 (not counting a perfect correlation with itself, of course). ALD is very close to both CALD and MEDO.

By computing complements to 1 of the correlation coefficients in **Table 3**, we obtain a distance matrix which can be used as input in hierarchical clustering. A cluster tree (dendrogram) from these data using the single-linkage approach is given in **Figure 7**. The dictionary branches connect at different levels, and the lower the linkage distance, the greater the connectedness.

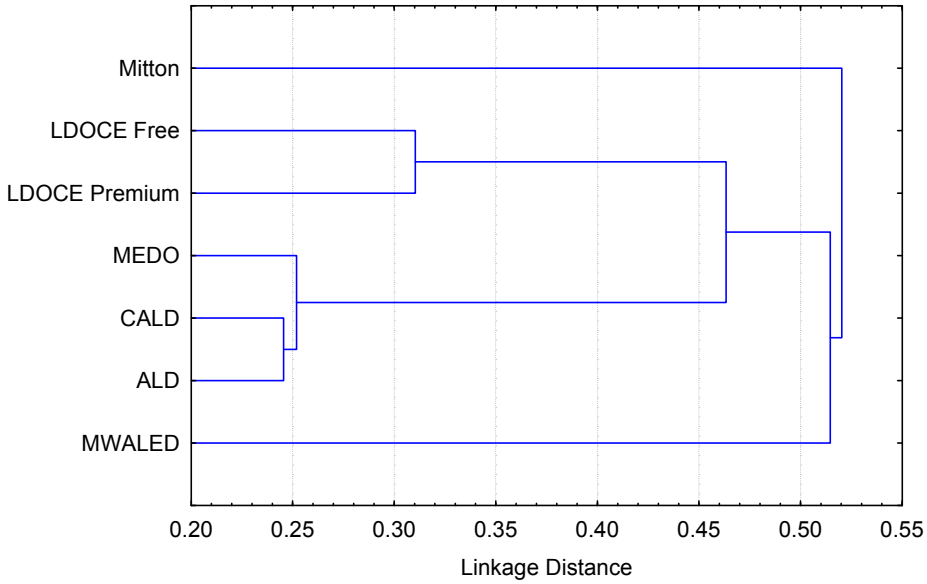

Figure 7: A cluster tree for the dictionaries on word rank data, using single linkage.

The dendrogram in **Figure 7** reveals a well-defined three-way cluster made up of ALD, CALD and MEDO, and another one comprising the two versions of LDOCE. These five dictionaries join together at the next step, bearing testimony to the common software platform. MWALED and Mitton's spellchecker remain relatively apart and increasingly distant from the core cluster.

6. Ways to improve spelling correction in e-dictionaries

6.1. Customization

While for many years the primary focus of research into spelling correction has been on native writers, recently the needs of non-native users of language speakers, particularly English, have begun to receive some attention (for an overview, see e.g. Heift and Rimrott 2008). It is now recognized that the patterns of misspelling of non-native speakers differ both in quality and quantity from those of native users of a language. Thus, if the L1 of the user is known to the system (be it based on the Accept-Language HTTP header, IP Geolocation, or individual user profile), the dictionary interface might use an algorithm optimized for that native language. In fact, Mitton's spellchecker used in this study has already seen a successful adaptation to better handle the typical misspellings of Japanese learners writing in English (Mitton and Okada 2007).

A non-trivial proportion of the items at which all the dictionaries failed are recognizable as attempts at rendering the pronunciation of the English word through the spelling conventions of the native language. This is particularly evident in the case of the Polish data, no doubt partially as a result of using audio stimuli for data elicitation. Evidence for this 'phonetic access' strategy (here largely subconscious, cf. Sobkowiak 1999) is seen in the use of L1-specific letter combinations (such as, for Polish, <sz>, <aj>, or <ej>) to approximate English pronunciation. Mitton's spellchecker handled many of these cases quite well, perhaps thanks to its level of phonological awareness, even though it has never been made aware of any Polish-specific letter-to-sound correspondences. Making provision for a few of the most common such correspondences could significantly improve a spellchecker's performance.

However, we would not expect the influence of L1 to be uniform across a wide range of L2 proficiency levels. To account for this variation as well as for individual idiosyncrasies, customization might in the future go even further: it might be possible to design an adaptive spelling corrector, capable of tuning in to the particular spelling problems exhibited by a given user, at a particular stage.

6.2. Dealing with real-word errors

In section 2.2. we discussed the issue of rare words. To use a specific example from the study, one of the misspellings in the corpus was **wold* for *would*. As it turns out, *wold* is also an English word, albeit very rare. Consequently, most occurrences of *wold* in text will be misspellings, and a text spellchecker would do well to flag it as a possible error. However, in a dictionary look-up situation, unlike in text spellchecking, it would be risky to withhold a rare-word entry from the user and offer instead similarly-spelled frequent words. Even though the core vocabulary of a few thousand words (De Schryver et al. 2006) are looked up more commonly than the rest, it is also true that the less frequent items have a reasonable chance of being looked up (see the discussion in 2.2. above). How should a dictionary respond to such a query?

The answer need not necessarily be the same for *any* dictionary. A user of the online version of, say, the OED is much more likely to want an entry for a relatively obscure word than a user of an intermediate-level learners' dictionary. The latter dictionary might not hold the word in its wordlist at all, in which case the issue would not arise. But if it did, a happy compromise might be to take the user to the rare word entry, but at the same time alert them in a sidebar saying something like 'Did you perhaps mean *world*'?

6.3. First things first

We have suggested possible avenues to improve success in correcting misspelled dictionary search terms. However, it needs to be stated emphatically that it would be misguided to pursue any such attempts at tweaking the interface before more basic problems are addressed. This study has revealed that such fundamental problems are numerous and grave, and they affect the most authoritative of English monolingual learners' dictionaries. Only a small minority of these problems have been addressed over a period of one year.

7. Conclusion

Our study has shown that the online versions of the leading monolingual English learners' dictionaries are inadequate when it comes to correcting misspelled input from non-native users. Far too often, when challenged with a misspelling, the dictionaries are unable to include the word actually intended in their list of suggestions, and, if they do include it, it often appears some way down the list. While the individual dictionaries vary substantially in performance, there is much room for improvement for even the best ones, and we have shown that an experimental spellchecker achieves much greater success rates than any of the dictionaries, even though it has not been designed with non-native speakers in mind.

8. Acknowledgements

The first author wishes to thank his student assistants, Marta Dąbrowska and Aleksandra Lasko, for their help in collecting the Polish corpus of misspellings.

Notes

[1] In terms of aims, approach and content, this article largely replicates Lew and Mitton (2011), which is not available in printed form. However, all the data are newly collected after a period of one year.
[2] http://grammar.about.com/od/words/a/misspelled200.htm
[3] For example, to get directly to the Google dictionary entry for the word *bay*, one would at this time use the following URL: http://www.google.com/search?q=bay&tbs=dfn:1. In some browsers (Opera, for example), it is possible to define customized search shortcuts of this type, so that lookups in the Google English Dictionary can still be performed conveniently from the address bar.

References

A. Online dictionaries tested

ALD. *Oxford Advanced Learner's Dictionary*. http://www.oxfordadvancedlearnersdictionary.com/

CALD. *Cambridge Advanced Learner's Dictionary*. http://dictionary.cambridge.org/

GoogleED. *Google English Dictionary*. At the time of collecting data: http://www.google.com/dictionary; at the time of writing the present version: http://www.google.com/search?q=%s&tbs=dfn:1 (where %s stands for the search term)

LDOCE Free. *Longman Dictionary of Contemporary English*. http://www.ldoceonline.com/

LDOCE Premium. *Longman Dictionary of Contemporary English*. http://www.longmandictionariesonline.com/

MEDO. *Macmillan English Dictionary Online*. http://www.macmillandictionary.com/

MWALED. *Merriam-Webster's English Learner's Online Dictionary*. http://www.learnersdictionary.com/

B. Other dictionaries

Hanks, Patrick and Judy Pearsall (eds.) 1998. *New Oxford Dictionary of English*. Oxford: Oxford University Press.

McKean, Erin (ed.) 2005. *New Oxford American Dictionary*, 2nd edition. Oxford: Oxford University Press.

C. Other literature

Atkins, Beryl T. Sue 1996. 'Bilingual Dictionaries - Past, Present and Future' in Gellerstam, Martin, Jerker Jarborg, Sven-Göran Malmgren, Kerstin Noren, Lena Rogström and Catarina Röjder Papmehl (eds.), *EURALEX '96 Proceedings*. Göteborg: Department of Swedish, Göteborg University, 515-546.

Bank, Christina 2010. *Die Usability Von Online-Wörterbüchern und Elektronischen Sprachportalen*. M.A. Thesis. Institut für Informationswissenschaft und Sprachtechnologie, Universität Hildesheim

Bogaards, Paul 2010. 'The Evolution of Learners' Dictionaries and *Merriam-Webster's Advanced Learner's English Dictionary*' in Kernerman, Ilan and Paul Bogaards (eds.), *English Learners' Dictionaries at the DSNA 2009*. Tel Aviv: K Dictionaries, 11-27.

Damerau, Fred J. 1964. 'A Technique for Computer Detection and Correction of Spelling Errors.' *Communications of the A.C.M.* 7: 171–176.

De Schryver, Gilles-Maurice, David Joffe, Pitta Joffe and Sarah Hillewaert 2006. 'Do Dictionary Users Really Look up Frequent Words? – on the Overestimation of the Value of Corpus-Based Lexicography.' *Lexikos* 16: 67-83.

Deorowicz, Sebastian and Marcin Ciura 2005. 'Correcting Spelling Errors by Modelling Their Causes.' *International Journal of Applied Mathematics and Computer Science* 15.2: 275-285.

Hanks, Patrick 2009. 'Review of Stephen J. Perrault (Ed.). 2008. *Merriam-Webster's Advanced Learner's English Dictionary*.' *International Journal of Lexicography* 22.3: 301-315.

Heift, Trude and Anne Rimrott 2008. 'Learner Responses to Corrective Feedback for Spelling Errors in CALL.' *System* 36.2.

Kukich, Karen 1992. 'Techniques for Automatically Correcting Words in Text.' *Computing Surveys* 24.4: 377-439.

Levenshtein, Vladimir 1966. 'Binary Codes Capable of Correcting Deletions, Insertions and Reversals.' *Soviet Physics - Doklady* 10.8: 707-710.

Lew, Robert 2011. 'Online Dictionaries of English' in Fuertes-Olivera, Pedro A. and Henning Bergenholtz (eds.), *E-Lexicography: The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum, 230-250.

Lew, Robert and Roger Mitton 2011. 'Not the Word I Wanted? How Online English Learners' Dictionaries Deal with Misspelled Words' in Kosem, Iztok and Karmen Kosem (eds.), *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10-12 November 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, 165-174.

Lindberg, Christine (ed.) 2006. *Oxford American College Dictionary*, 2nd edition. Oxford: Oxford University Press.

Mitton, Roger 1985. Birkbeck Spelling Error Corpus.

Mitton, Roger 1996. *English Spelling and the Computer*. Harlow: Longman.

Mitton, Roger 2009. 'Ordering the Suggestions of a Spellchecker without Using Context.' *Natural Language Engineering* 15: 173-192.

Mitton, Roger and Takeshi Okada 2007. The Adaptation of an English Spellchecker for Japanese Writers. *Symposium on Second Language Writing*. Nagoya, Japan.

Nielsen, Sandro 1995. 'Alphabetic Macrostructure' in Bergenholtz, Henning and Sven Tarp (eds.), *Manual of Specialised Lexicography*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 190-195.

Okada, Takeshi 2005. 'A Corpus-Based Study of Spelling Errors of Japanese EFL Writers with Reference to Errors Occurring in Word-Initial and Word-Final Positions' in Cook, Vivian and Benedetta Bassetti (eds.), *Second Language Writing Systems*. Clevedon: Multilingual Matters, 164-183.

Pollock, Joseph J. and Antonio Zamora 1984. 'Automatic Spelling Correction in Scientific and Scholarly Text.' *Communications of the A.C.M.* 27.4: 358–368.

Proctor, Edward 2002. 'Spelling and Searching the Internet: An Overlooked Problem.' *The Journal of Academic Librarianship* 28.5: 297-305.

Sobkowiak, Włodzimierz 1999. *Pronunciation in EFL Machine-Readable Dictionaries*. Poznań: Motivex.

Suomi, Riitta 1984. *Spelling Errors and Interference Errors in English Made by Finns and Swedish-Speaking Finns in the 9th Form of Comprehensive School*. MA Thesis. Department of English, Abo Akademi, Abo.

Véronis, Jean 1988. 'Computerized Correction of Phonographic Errors.' *Computers and the Humanities* 22: 43-56.

Wagner, Robert A. and Michael J. Fischer 1974. 'The String-to-String Correction Problem.' *Journal of the Association for Computing Machinery* 21.1: 168-173.

Yannakoudakis, Emmanuel J. and David Fawthrop 1983. 'The Rules of Spelling Errors.' *Information Processing and Management* 19.2: 87-99.

Appendix: Sample misspellings

SUBCORPUS	TARGET	MISSPELLING
PL	certain	serten
PL	easily	izli
PL	guarantee	garanti
PL	interfere	interfir
PL	interruption	interapsion
PL	library	lajbery
PL	psychology	sakology
PL	receive	reseve
PL	separate	sepret
PL	succeed	sukcid
JP	albatross	albatlos
JP	antenna	untena
JP	beautiful	butiful
JP	embarrass	enbarance
JP	enough	inaf
JP	gallery	garally
JP	graph	glaf
JP	laughter	lafter
JP	neglect	nigrect
JP	umbrella	umblera
FI	because	becourse
FI	colour	coulor
FI	delicious	delecous
FI	especially	espessially
FI	gasoline	gazolin
FI	good-bye	goodbay
FI	orchestra	orkester
FI	symphony	sinfony
FI	temperature	tempeture
FI	universities	univercitys