*TALŌ B.V.
the Germanic root of our words *tell*, *tale*, like a *notch made on a tally*

Lijsterlaan 379
1403 AZ Bussum
The Netherlands
http://www.talo.nl

tel +31 35 69 32 801
fax +31 35 69 75 993
cell +31 65 46 83 544
e-mail jaapw@talo.nl

Bank: Rabobank 39.47.33.339
IBAN: NL90 RABO 0394 7333 39
Swift-address Rabobank: RABO NL 2U
Handelsregister Gooiland 32045548

The *TALŌ Spell Checker, a high profile tool.

Orthography is the art of writing words with the proper letters according to the standard usage; the correct spelling agreed by language institutes opposed to ***cacography***, bad handwriting, or in modern times, a bad system of spelling.

Most spellcheckers suffer from poor spelling algorithms, even the English language is affected by this phenomenon.
Let's discuss the subject ***cacography*** in British English and focus on orthographic errors unnoticed or improperly processed by many spellcheckers.
What is the source of the problem and why do they not replace these poor algorithms?

The source of most of orthographic mismatches comes from a very loose permutation technique introduced during the 80's of the last century. Such a permutation technique was introduced as an attempt to increase the size of word lists by artificial means. At first glance permutations did create an extremely large virtual lexicon, and attention was only focussed on bigger, bigger, bigger. However, the fairy tale proved to be less effective in the end. Some of the permutations were correct, but most of them just didn't even exist. Usually a lot of effort was taken to define constraints in order to reduce the infinity of combinations known from mathematics (calculate x! (x faculty) for x=1000 on your calculator).

The poor performance of the permutation algorithm calls for a more accurate analysis. A virtual lexicon which consists of three-quarters of errors does not recognize three-quarters of the mistakes people make. Most of the errors in publications came from unthinkingly accepting the speller's results. So the assumption that permutations of root words can replace a full-size lexicon is not TRUE. Moreover, the leakage of weird combinations into the spellchecker's alternative list reduces its useability. In quite a few cases people even set these spellers aside.
A second obstacle arises from the misuse of hyphens and spaces in compounded words (***well-being, well-thought-out, well enough, bail-bond, bail bondsman, bailout, to bail out, etc***). It has an immense influence on the performance of nearly all spellcheckers unless provisions are given. The presence or absence of hyphens or spaces affects the meaning of compounded words significantly. A blue book is a book bound in blue, but a ***grand master*** is the chief of a royal household and a ***grandmaster*** usually denotes to a chess player of the highest class.
The above holds true for many languages.

A British English example of a ***cacographical*** text demonstrates the capability of spellcheckers, and it shows what is often seen and what has to be flagged by a spellchecker. Below the sections to be flagged are printed in a slightly larger font.

> *The* *hot-favourite* wishes to be a **competiter**† and rushes *hot-footed* to the planning office to object. *Hotheaded* as the person was he entered the hothouse with a hot key and put his name on the hot list. Being responsible he took his hot seat, he just had a bad week and was bad-tempered. There it was, the bad news. He thought no one wants to hire an individual who had bad-mouthed *an prior* employer. There are two *13th century* **Islandic**† books **havin**† a collection of legends, from *great-grand-mother's* time.
>
> I'm not bragging about being a **wreckless**† *driver*, I know I *drive* **wrecklessly**†. His *ship was* **recked**‡, and he was drowned. The only living creatures were **reckfishes** in the warm **atlantic**† waters. In a **recking**† *ball* hung.

---

*TALŌ is the Germanic root of our words tell, tale, and tally.
It also is the mark of linguistic software which has its roots in neurobiology and human factors.

The morning star was a *write off*, its credibility rating below zero the morning after. *Morning after pill* did not make a difference to the **morfofonological** nature of the morris dancer around his *morris chair*, its arm was **splitted**, its *wooden head* burned of stupidity when the *wooden top* was addressing the subject of **advertizing**† to him. *Ad-hoc* as his plan was his *wood carving* skill was missing.

The launching of a spacecraft to the moon is called a *moon shot*, but *some times* it looks like a **moon-struck** activity.

As human misstakes cross the word boundary the example *moon shot* in the last sentence should be flagged as an entity. It is not enough to approve the word **moon** and thereafter the word **shot**.

A well-known American spellchecker we tested had 8 hits of the 28 mistakes only. These hits are marked in bold and they got an additional dagger † if successful (the correct suggestion). A double dagger ‡ signals a false alarm (2x), e.g. both wrecked and recked exist and a mistake can only be solved in context. If a hyphen was inserted incorrectly (e.g. ad-hoc) the mistake was just skipped. A mistake in an expression (a space) was never recognized. Moreover suggestions were quite often dissimilar from the mistake, see table 1.
In addition to bold marked mistakes the enlarged sections were recognized by *TALO's Speller (errors within context, e.g. a ship in relation to a wreck), see also Annex A.

Table 1.: Suggestions shown by four different spellcheckers, TALO's spellchecker has a hit, but App OF spellchecker's results are poor, APP MW only shows one look-alike, but no hit, and App ID spellchecker has problems with some plurals derived from Latin. Both App OF and ID send suggestions not really similar to some extent.

spelling mismatch:
reckfishes

got alternatives from TALO:
wreckfishes (hit)
rockfishes

got alternatives from App OF:
refinishes (ends on ....ishes)
refurbishes (ends on ....ishes)
cuttlefishes (ends on ....ishes)
crawfish's
breakfaster

got alternatives from App MW:
rockfishes

got alternatives from App ID:
reck fishes (not different)
rockfishes
weakfishes
rachises (error rachis -> pl rachides)
rockfish
ricochets
reacquires
rake-offs
reequips (should be re-equips)
requiems

Re-spelling between British and American orthography.

*TALO's Speller also includes a re-spelling mechanism. This mechanism recognizes Americanisms in British English texts and proposes a British English Alternative, e.g.,

"The **colors** that have been selected ..." >> "The **colours** that have been selected ...
"There is the woman whose technical report won top **honors** ..." >> "There is the woman whose technical report won top **honours** ..."
"The match was **canceled**" >> "The match was **cancelled**"
"We **traveled** to Paris" >> "We **travelled** to Paris"

The above corrections can be executed automatically or the user has to approve (see Annex B fig. 1.).
The reverse — British to American English is also possible.

This re-spelling mechanism is also used for more complex cases such as:
"**Lloyds** register" >> "**Lloyd's** Register"
"**Lloyd's** Morgan's canon" >> "**Lloyd** Morgan's canon"

to recognize a difference between **LLoyd's** (insurance incorporated association) and **Lloyd** (Brit. Psycholo-

gist).

These re-spelling mechanisms are essential for languages whose orthographies were reformed (German, Dutch, Portuguese, or French recommended).

British English and American English differences

In both versions of the English language hyphens and spaces present a dilemma. As was demonstrated above traditional spellcheckers just don't see many mistakes. This has to do with a poor procedure to analyse mistakes. For instance: *Taxpayer* is a closed compound and correct in both British English and American English, unlike **tax payer** or **tax-payer**. An **afterdinner cup** is an error too, it should be *after-dinner cup* but if afternoon is used as a noun it has to be *afterdinner*. While in American English *makeup* is correct, it is *make-up* in British English. A *Druid stone* might be found at Stonehenge, but there will be neither a **Druid-stone** nor a **Druidstone**. An altogether different type of word is an eggcorn[1], which sounds like acorn. An eggcorn sounds like a legitimate word but isn't. **Soaping-wet** (as in **a soaping-wet T-shirt**) is an example of this category. It has to be a *sopping-wet T-shirt*. **Soaping-wet** sounds like *sopping-wet*. Two other examples are: **coal-hearted** (*cold-hearted*) and **coal-blooded** (*cold-blooded*). The reason why these odd words are not noticed by many spell checkers is technically related to the erroneous way of processing hyphens and spaces.

An alternative

As stated above most spelling checking programs cannot deal adequately with hyphens and spaces because they are based on the derivation of words from elementary roots. The results are very often words that do not really exist and, consequently, errors are not recognized, as is apparent from the examples shown in this article. Instead of a derivation of words from elementary roots an appropriate language-specific model is required, i.e. a language model that includes descriptions of characteristics of an individual language. This is the approach taken by *TALŌ[2], to be able to deal successfully with the *"hyphen / no hyphen / space / no space"* confusion as demonstrated in this article. An external speller library is available to customize a spellchecker.

Jaap Woestenburg PhD, Bussum

Annex A: Erroneous and Correct Orthography

hot-favourite >> hot favourite (a noun and not an adjective!)
competiter >> competitor
hot-footed >> hotfooted
hotheaded >> hot-headed, but hothead (noun) is without a hyphen
an prior >> a prior
13th century >> 13th-century (attributive usage only)
Islandic -> Icelandic
havin >> having
great-grand-mother's >> great-grandmother's
wreckless driver >> reckless driver
drive wrecklessly >> drive recklessly
ship was recked >> ship was wrecked (reck as a verb exists(‡), but a ship can only be wrecked, see fig. 2)
reckfishes >> wreckfishes
recking ball >> wrecking ball (reck as verb exists but in context a wrecking ball is the only possibility)
a write off >> a write-off
morning after pill >> morning-after pill
morfofonological >> morphonological
morris chair >> Morris chair, but morris dancer is in lower case
was splitted >> was split (the form splitted has disappeared in modern English)
wooden head >> wooden-head
wooden top >> woodentop
advertizing >> advertising

[1] "http://eggcorns.lascribe.net/"

[2] "http://www.talo.nl/"

ad-hoc >> ad hoc
wood carving >> woodcarving
moon shot >> moonshot
some times >> sometimes
moon-struck >> moonstruck

EXPRESSIONS or COLLOCATIONS
COLLOCATIONS (with LL) are word sequences which are part of a more complex lexicon. The collocation (multiple word) lexicon focuses on user errors, the correct usage itself is a waste of storage and is never used.

Most principles apply to other languages too.

Annex B: Images

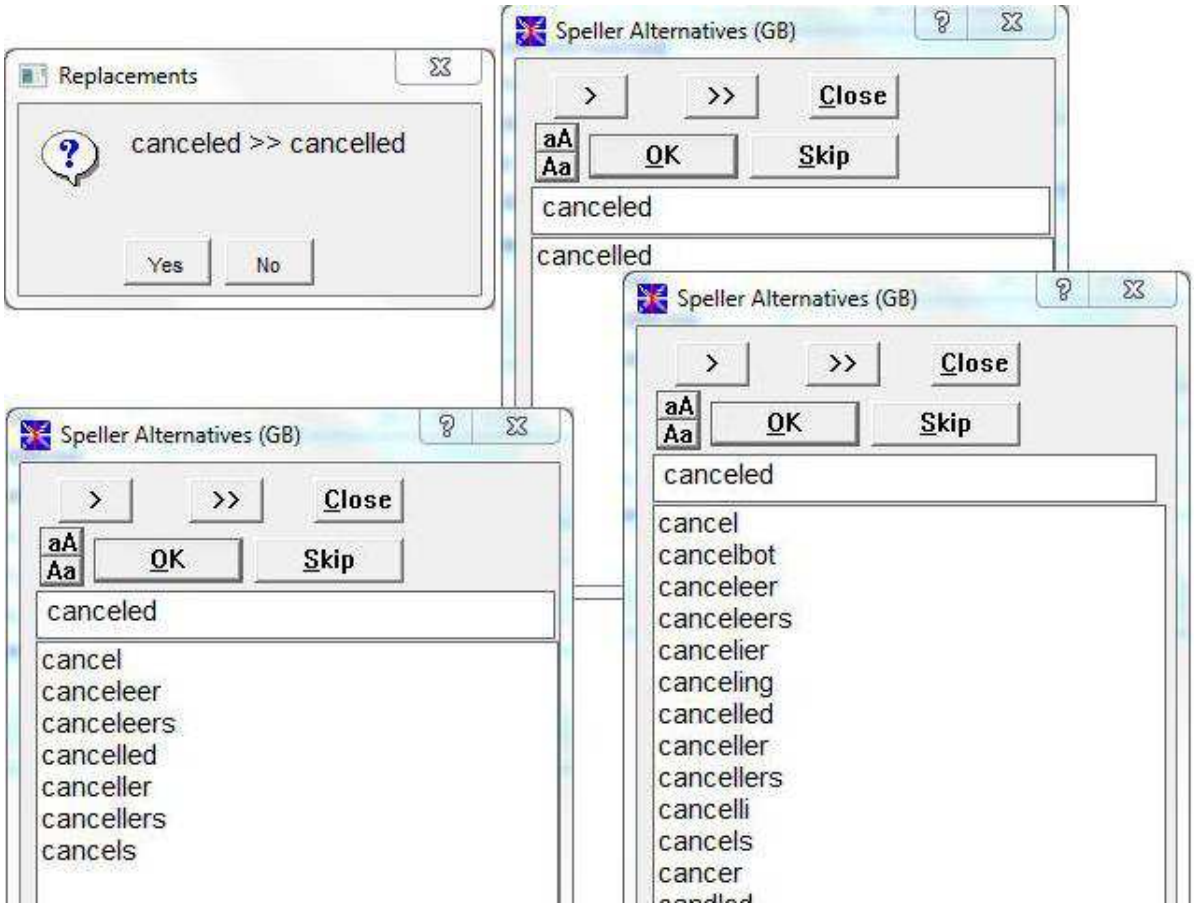

Fig. 1.: Respelling from American to British orthography, replacements redirected to the dialog, request more [>] alternatives, request most [>>] alternatives.

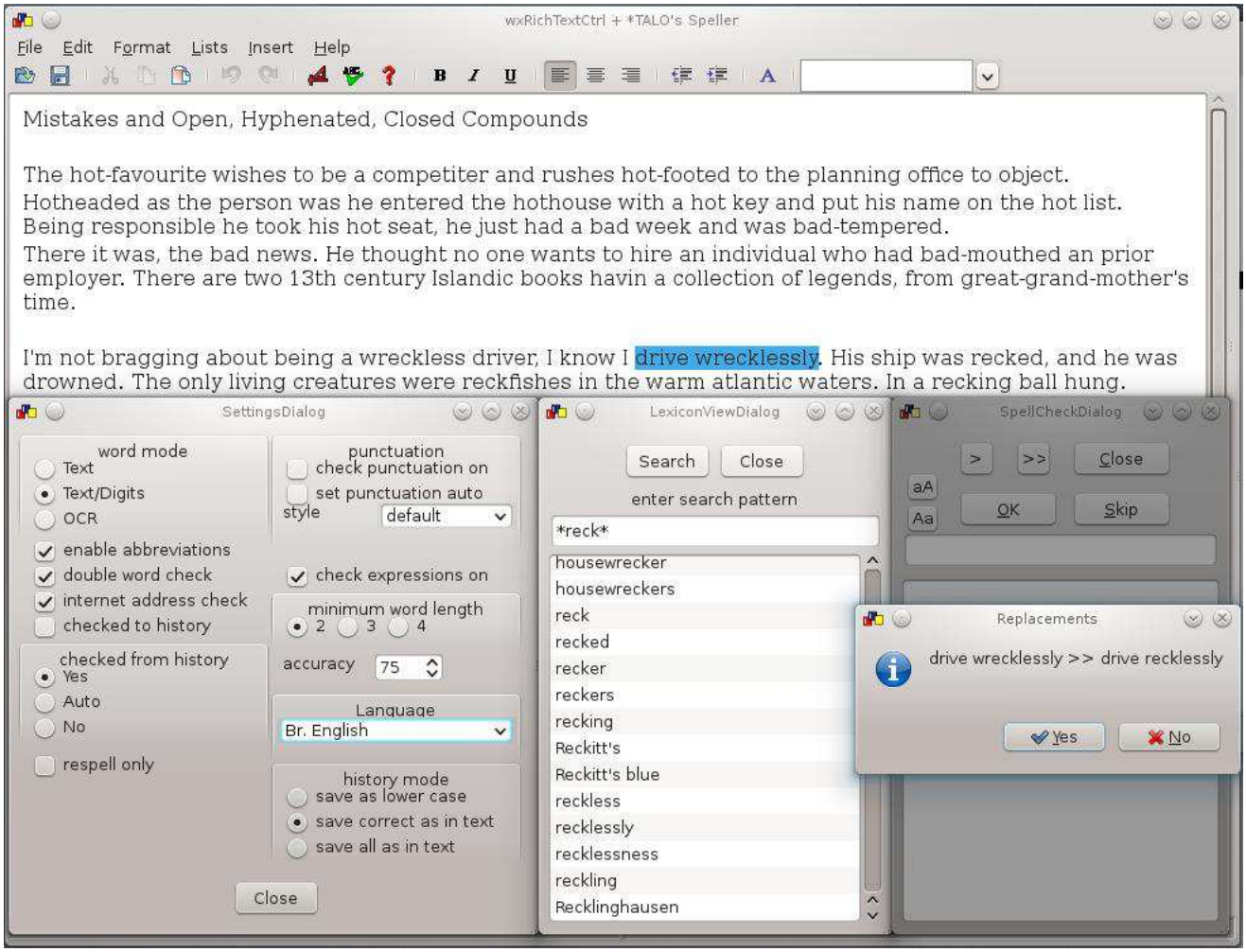

*Fig. 2.: RichText Tool showing the Spellers's SettingsDialog, the LexiconViewDialog and the SpellCheckDialog. The language has been set to British English. There is a search for the pattern "*reck*", showing a few of the (w)reck cases in the lexicon and the correction in context is presented (Replacements). The LexiconViewDialog and SpellCheckDialog are resizeable.*