*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

*TALŌ's Language Technology 6.2.3.7.
Make-or-Break Your English

TALO's language models have been renewed. Learning corpora have been enlarged and tuned to the best performance ever. Support of special hyphenations have been renewed for several languages, e.g. the Finnish language. The existence of these features of language are unknown to many IT-system developers. Still they are necessary to process language properly.

The new language models have been integrated into the spellcheckers too. These models increase the power to suggest the most lookalike alternatives. Sometimes users hit incorrect keys, or just omit a letter. Still the majority of these mistakes can to be solved and, of course, lexicons have grown extensively.

Blind and bewildered

Some spellcheckers are known to blindly accept a great deal of combinations in many languages. Most of these combinations are certainly wrong. Why do people accept these omissions? Their self-esteem? If your text is full of orthographical errors, does it represent your language skills? Maybe it is not so bad, you only trusted your spellchecker too much.

American and British orthography is different, and it is not only a matter of "harbor" versus "harbour". Most trusted American dictionary publishers (The Unabridged Meriam Webster) present us with the American view on orthography. However, the Oxford publishers deviate from this view (the shorter Oxford English having two bands only instead of the full-sized 13 band). The main topic is the possibilities of building English compounds, words which consist of two or more independent composites.

An important class of nouns

A front runner (O) | front-runner (W) | frontrunner
A fund raiser | fund-raiser (W, O) | fundraiser
A game bird (W,O) | game-bird | gamebird
A kick starter | kick-starter (W,O)| kickstarter
A girl friend (W)| girl-friend | girlfriend (O)
Trend setters | trend-setters | trendsetters (W,O)
A cross bencher | cross-bencher (O) | crossbencher (W)
(0) = Oxford, (W) = Webster

Broken promises of a kick-starter

MS word (et al.) only considers "kickstarter" as wrong and suggests you try "kick-starter", "kick starter", or "kicks tarter" (only kick-starter is correct). MS orthography of girlfriend doesn't agree with the Unabridged Meriam Webster's. In the case of attributive usage open compounds are hyphenated (front runner (O) -> front-runner horse.

A few questions arise: a) how does MS (et al.) evaluate orthography?, and b) how to deal with noun and attributive cases of compounds? The creativity of newspapermen demonstrates how MS spellchecks the word "Disney-fication". MS only marks "fication" as an error which is indeed horrible. MS does not red-line the whole word, only the section after the hyphen is red-lined. In fact, the hyphen should have been left out "Disneyfication". The word itself is a more British than American. However, it is true that MS accepts nearly all possibilities of openness (a space), hyphenatedness (a hyphen), or closedness (nothing in-between). And, therefore, several severe mistakes remain unnoticed because the elements of the compound are correctly written.

To join or divide

We will demonstrate a few possibilities implemented in our speller design: multiple word correction. The most

*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.

important topic is: the hyphen in a word is very informative, but sometimes an extended context is required, e.g. the next word.

The most simple case to solve errors is the plurality of words, e.g.,

UK: front-runners -> front runners, a plural noun case (and not an attribute before a noun).

Other cases require context to differentiate between a noun and an attribute, e.g.,

UK/US: free soil (noun) -> free-soil (attribute), the Free-Soil party,
US/UK: stock exchange (noun) -> stock-exchange (attribute), never "stockexchange",
use "stock exchanges", "stock-exchange news", "stock-exchange organisation", et cetera.

Back to grammar school
Another grammar-a-like example of multiple word correction is:

> "will have seeking" -> "will have sought"

A sentence as "Her remarks were well intentioned" (adverb case) or "She had a well-intentioned remark" (adjective case) results in a very common mistake: "Her remarks were well-intentioned", which is grammatically incorrect. The multiple word correction would be:

> "were well-intentioned" -> "were well intentioned" (adverb + past participle)

The above correction has been implemented more efficiently to process all kinds of lookalikes, which seem to be a regular hit.

The English indefinite article "a" is a source of mistakes too. You have to add a "n" if a word starts with a vowel, but people often do this incorrectly: *an house, a effort*, or as was seen recently in the NYT and Guardian:

> "the battle took on *an striking* degree of partisanship." which ought to be "a striking ...",
> "The Givenchy show drew *an celebrity-filled* front row."
> "The murders on *a idyllic* beach shocked ...".

It is not always a written vowel, but a vowel sound, that is significant, e.g. in "an MP" (Member of Parliament) and a PM (Prime Minister).

The conclusion is that quite a few cases are ambivalent, either an attribute or a noun case is meant, but it might be worse. To solve these orthographic mistakes additional context is needed. If a word is only defined as a succession of alphanumeric symbols, including apostrophes and hyphens, it is not possible to solve the above ambivalence. However, if context is added, the spellchecker has to be capable of jumping over spaces too.
So "*blockchain technology*", has to become "*block-chain technology*", which is derived from the noun "block chain" (W). The selection consists of two succeeding words (each between spaces) instead of spellchecking the two words separately.

Uncertainty, blendings!
If erroneous orthographies are used in succeeding words, these mistakes have to be displayed on one line in a speller dialog. In addition, multiple word suggestions have to be sent to the dialog too.

> "pin a finger" (the mistake, a linguistic blending)
>
> "pin the blame on" (alternative 1)
> "point a finger at" (alternative 2)

To be able to do so, it is necessary to spellcheck paragraphs instead of words and look out for known multiple word mistakes. The choice of dialogs is either as compact as possible, or a composite of many buttons. It is our personal choice to use dialogs in demos which are as compact as possible, however, we offer developers functions to build their own dialogs keeping their own corporate style. Multiple word spellchecking can be implemented independently of any dialog style.

Not only writing well, but also spellchecking well, is **a make-or-break skill** (BBC 17 sept. 2015).

For additional information go to: *http://www.talo.nl/* Bussum, 27 October 2015