*TALŌ's LANGUAGE TECHNOLOGY
A Commentary on Spelling

Dr.J.C.Woestenburg,

**TALO b.v.,*
Lijsterlaan 379,
1403 AZ Bussum,
The Netherlands.

tel: +31 35 69 32 801; fax: +31 35 69 75 993
e-mail: info@talo.nl; http://www.talo.nl/

Bussum, August 2004

ABSTRACT

*TALO has developed new lexicon technologies that optimize content as well as accessibility. These technologies even go beyond the word boundary and handle combinations of consecutive words, abbreviations, punctuation and even style. This results in more efficient and more effective spell checking. The importance and the relevance of this new approach are discussed in this article.

Compounds appear in most languages; people create them every day. Unfortunately, when these extended words are checked by currently used spellers many words spelled incorrectly are nevertheless accepted as having been spelled correctly. These spellers are based on outmoded technologies and users often are not aware of the poor performance of such outmoded spellers.
This article reviews the traps a user could fall into and provides quantitative results of performance comparisons between different spellers including *TALO speller.
It is beyond the reach of many spellers to include in their lexicons all of the compounds that appear in a particular language. That is why their lexicons consist mostly of the roots of words. These root words are meant to be combined with one another. But such combinations, also called permutations, are applied blindly, which leads to the acceptance of spelling errors without the user's approval.
These spellers may also provide suggestions that are plainly absurd.
On the other hand, research has shown that the number of irrelevant alerts (bad flags) decreases with increasing size of the lexicon. This emphasizes the importance of a large, relevant lexicon.
Research has also shown that this increase has little or no effect on the number of missed errors.

Currently used spellers may be based on the permutation method or the trigram method. Both methods miss errors in compounds (or, in other words, they approve errors). That is why besides the appropriate technology the focus should be on building large lexicons with a large variety of information. These modern lexicons cover a substantial segment of national idioms.

Sometimes a national spelling revision surpasses the old technology. For instance, the German spelling revision has been expanded to cover groups of words that are used in more or less specific contexts but are recognizable with *TALO's new technology.

SAMENVATTING

*TALO heeft nieuwe woordenboektechnologieën ontwikkeld die inhoud en toegankelijkheid optimaliseren. Deze technologieën gaan zelfs voorbij de woordgrens en behandelen combinaties van opeenvolgende woorden, afkortingen, interpunctie en zelfs stijl. Dit resulteert in efficiëntere en effectievere spellingcontrole. Het belang en de relevantie van deze nieuwe benadering worden besproken in dit artikel.

Samenstellingen komen in de meeste talen voor, men creëert ze dagelijks. Wanneer deze lange woorden gecontroleerd worden door gangbare spellers worden veel verkeerd gespelde woorden toch geaccepteerd als zijnde goed gespeld. Deze spellers zijn gebaseerd op achterhaalde technieken en gebruikers zijn zich vaak niet bewust van de slechte prestaties van zulke achterhaalde spellers.
Dit artikel laat de valkuilen de revue passeren waarin een gebruiker zou kunnen vallen en geeft kwantitatieve resultaten van vergelijkingen tussen verschillende spellers, inclusief *TALO's speller.

Het ligt buiten het bereik van veel spellers alle samenstellingen, die in een bepaalde taal voorkomen, op te nemen in hun woordenlijsten. Daarom bestaan hun woordenlijsten veelal uit stammen van woorden. Het is de bedoeling dat deze stammen met elkaar gecombineerd worden. Zulke combinaties, ook wel permutaties genoemd, worden echter blindelings toegepast, wat leidt tot de acceptatie van spelfouten zonder goedkeuring door de gebruiker. Deze spellers doen ook weleens suggesties die kant noch wal raken. Aan de andere kant blijkt uit onderzoek dat er minder irrelevante alarmeringen (bad flags) zijn naarmate het lexicon groter is. dit onderstreept het belang van een groot, relevant lexicon. Onderzoek heeft ook aangetoond dat deze toename in omvang weinig of geen invloed heeft op het aantal gemiste fouten (missed errors).

Gangbare spellers kunnen gebaseerd zijn op de permutatiemethode of de trigrammethode. Beide methoden missen fouten in samenstellingen (of, anders gezegd, ze keuren fouten goed). Daarom

moet, naast de juiste technologie, de aandacht gericht zijn op het bouwen van grote lexicons met een grote variëteit aan informatie. Deze moderne lexicons dekken een groot segment van nationale idiomen.

Soms overtreft een spellinghervorming de techniek in bestaande spellers. Zo heeft, bijvoorbeeld, de Duitse spellinghervorming zich uitgebreid naar woordgroepen die min of meer in specifiek verband gebruikt worden, maar met *TALO's nieuwe technologie herkenbaar zijn.

NEW SPELLING TECHNOLOGY

The *TALŌ spellers[1] use new technologies to overcome the common problems found in so many spelling tools. Many tools use wrong assumptions based on the wrong language. They only cover a language idiom rudimentary, try to camouflage this shortcoming by easily accepting the unknown, and if not they often offer unrelated suggestions. This calls for a new approach to spelling.
These new technologies include features such as:
a) unfolding lexicons with a wide variety of words including conjugations and inflections, belonging to the real stock of words. This means that a large number of words of all types is stored in the lexicon and is instantaneously accessible.
b) making the full linguistic information available by leaping up and down. This means that words in the lexicon are not searched by a linear method, but in jumps right to the spot where the information is. Information is available without delay.
c) making use of very specific language models each tuned into one target language. The model knows the relations between words and looks for the perfect word structure to find related suggestions in case of a spelling error.
d) making use of new comparisons to estimate the amount of similarity, just on the fly. These comparisons keep the syllables, lemmas, and other morphological information in the right order.
e) compact information using linguistics, instead of compression.

The new technologies do away with a lot of risky operations in spelling. There is no need to permute words (combinations), because the proper cases are already in the lexicon. The advantage is that the meaningless combinations never occur. All exceptions of intuitive rules are available to the speller engine.

New technologies have been developed to handle difficulties situated outside the traditional spelling of words: abbreviations, punctuations, combinations of words. Therefore the spellers do not spell per word, but test orthography in larger units, a sentence, a paragraph, or even paragraph after paragraph. The advantage is that each kind of error is tested with its own optimized algorithm.

New technologies have been built in to let the speller learn (to acquire knowledge by experience to be useable later on). It also could be described as a learning system that rehearses or forgets. It is very different from the usual speller's Add and Delete button. The latter functionality only modifies a list in terms of adding or deleting a record, without learning from previous user failures in the text. For the new technologies the net effect is that every error in words will be evaluated with regard to relevance. The functionality of the concept of accuracy

helps the user to retrieve the proper word, or to maintain a style often presented in style guides.

These new technologies are based on research and are applied to over 70 languages or varieties. This variety guarantees that we have seen all peculiarities in languages. Consequences have been considered ahead of time before they are applied in a professional setting. Despite these varieties in language our tools are presented as a single uniform method accessed by the user.

The performance of these new technologies differ considerably from the performance of earlier speller technologies. Therefore the performance of these earlier technologies has to be analysed in detail.

WHAT ABOUT EARLIER SPELLERS OR EARLIER TECHNOLOGIES

Prior to discussing the speller mechanisms and their limits we should first state "what performance do we expect from a speller"! A speller should detect any error in the text. The word "any" might be an Utopian requirement, but the ideal speller's performance should approach this requirement as closely as possible.

People who create texts make several types of errors:

a) errors in single words ("hause" for house),
b) errors in a word that depends on the context "an house".
c) abbreviation errors (23-mm instead of 23mm).
d) punctuation errors (,,a typewriter's citation'').

If an error occurs a speller should send a warning message! Automatic correction might be possible but language is quite complex and a warning message is to be preferred. However, spellers get into trouble, because their lexicons are too small. Moreover the word itself might be correct, but word combinations might be wrong. A lexicon should cover a very large section of the real population of words (the total collection of words that are used by an extensive group of people). This means that errors can only be tested on their merits using large lexicons, including provisions to detect wrong combinations.

The very first question is: how did spelling get started? Even now, old technologies are still in use. Text is compared with simple text files of single words, ordered alphabetically. Unix systems have the old AT&T spell function, a plain word list for the English language. The list itself is too small to be useful. Such a list matches only a fraction of the current idiom of a language. Usually only En-

glish and a home-brew list at the institute itself are available. Other spelling tools that came from Unix environments are ispell, aspell and a lot of variations on the same theme. Most of these tools originated from the North American universities and companies, focussing on the English language. These technologies are limited in scope, they are outdated.

Spellers in professional applications such as Microsoft's Word, Apple, Quark's QuarkXPress and Adobe's InDesign do NOT add more performance than free domain tools.

The question is: which mechanisms do spellers use and what are the drawbacks of these mechanisms.

The most important reason that calls for large dictionaries can be found in language itself. Language consists of clusters: the phonemes which make syllables, and the syllables which make words, and most frequently words are inflected (nouns) and conjugated (verbs). The building of blocks of words continues. Based on meaning a number of words belong together and these words form compounds. These compounds become a major feature of a language and the number of possibilities are numerous. Yet the way compounds have been and are being created is subjected to rules strictly dominated by meaning. Sometimes the foundation of meaning occurred in the past, but most of the new forms arise from new phenomenons in society.

For many spellers large lexicons are an Utopia, far out of reach, so they need to fall back on tricks.

If real word lists are small, pretty soon any word might look like an error (a mismatch). These mismatches would imply time consuming stops. For many people these stops are an argument of not using these spellers. These stops are the main reason to permute word roots and test each permutation against the error as shown below. If there is a match the user is not informed about the artificial nature to approve the combination, even if the combination was highly unlikely to be correct.

PERMUTATION, A TRICK

One of the tricks to disguise failures is to compose a lexicon of lemmas only. These lemmas are root words. Root words are permuted with each other, and all possible compounds are blindly assembled by an algorithm, independently

of the meaning of a word(s). The result could be a new word list, but such a word list will never be shown. The reason is simple. Let's start with the 2 words "fighter" and "knife" and make any possible combinations:

fighter knife
fighter knives
knife fighter
knife fighters

Example 1

These open (with a space) compounds sound reasonable in terms of meaning, but this is not a general rule. The two words "knife" and "sheath" do not nicely permute. "sheath knife" would be a combination without any meaning. This illustrates that the order of words is strictly tied to meaning.
The English language has many open compounds, words separated by a space, opposed to words written together. However, this English feature deviates from most languages in terms of openness. Most languages build closed compounds.

Danish: aktivitets|bestemt
Dutch: afbetalings|gedrag
Finnish: liike|voitto|prosentti
German: Abonnements|fern|sehen

Example 2

Example 2 shows that compounds can also include a genitive ending (or an s-binding, pronounced as ES-...).

To show that a blind permutation can be meaningless, the Danish "bestemts|activitet" would be a nonsense term, and so would be the Dutch "gedrags|afbetaling". Another Danish example:

aktivitetsbestemt
aktivitetbestemt
bestemtaktivitet
bestemtsaktivitet

Example 3

only the first one is meaningful while the other three words are nonsense words. Therefore the probability of a correctly generated compound is only 25% in this particular case.

In Finnish "jumalan|ilma" would be quite different from "jumala|ilma" (god-weather does not exist but god's weather does, and it would be rather stormy). Finnish is a highly inflected language with at least 14 different forms for the singular and 14 forms for the plural, and a lot of additional suffixes and clitics, but word cases are usually restricted to nominative and genitive.

The way words permute is different from language to language. For the Dutch language an s-binding, e-binding, en-binding, a hyphen-binding, or no letter in between at all exist.

dochters|goed,
gedachte|streep,
boeren|bruin,
documentaire-theater
chromaat|geel,

Example 4

The bindings are irregular and differences in meaning do have different bindings.
Not being aware of meaning a simple permutation of the Dutch words 'tulp' and 'manie' or 'gelofte' and 'dag' would result in:

tulpmanie (non)	geloftedag
tulpemanie (e-binding)	(is already an e-binding)
tulpenmanie (en-binding)	geloftendag
tulpsmanie (s-binding)	geloftsdag
tulp-manie (hyphen binding)	gelofte-dag
tulp manie (open compound)	gelofte dag
and the reverse cases	
manietulp	daggelofte
	dagegelofte
manientulp	dagengelofte
maniestulp	dagsgelofte
manie-tulp	dag-gelofte
manie tulp	dag gelofte

Example 5

However none of the tulp words is correct and only one of the gelofte words is correct. The gelofte cases have a probability of 1 to 12 of being correct. This technique is very likely to accept erroneous permutations.

So permutation is not the technology to be used in spelling. But what occurs if a speller behaves in such a way? How does one recognize the use of permutations in a speller?
Developers, who use such an algorithm, claim unrealisticly large dictionaries. However, the small size of the lexicon they ship does not reflect a large number of words.

What happens with these erroneous combinations is that, if they occur in the user's texts, and they do, they are likely to be approved as correct!

in Dutch:
tulpen|manie (should be tulpomanie),
kostwinnaar (should be kostwinner)
in German:
Asche|mittwoch (should be Aschermittwoch),
Empfangantenne (should be Empfangsantenne)

Example 6

The number of false approvals increases with increasing number of permutations. Some incorrect permutations can be avoided by adding morphological information to lemmas, a certain type of words always has an s-binding). However, most varieties are determined by the meaning of words (Dutch: registreerapparaat and not registratieaparaat).

Assuming two lemmas only, chances are between 3/4th and 1/9th of the possibilities would be incorrect. The more lemma's the more combinations are possible and a very few would make sense. Fortunately some errors are unlikely to be made by people. Yet there is a substantial probability that erroneous combinations occur and are approved.

These errors appear in quality newspapers, quite often related to the journalist's misunderstanding of words:

Dutch:
oudjaarsdag, flagellaten, incorpereren, islamistisch, marsepijn
instead of the correct
oudejaarsdag, flagellanten, incorporeren, islamitisch, marsepein
(in English: New Year's Eve, flagellants, incorporate, Islamic, marzipan)

Example 7

The conclusion we can draw is that permuting is a useless, risky feature that lets the user accept mistakes.

WHAT ABOUT PERFORMANCE?

A key aspect to be discussed is whether methods exist to prevent the recognition of all possible errors and on the other hand to reduce the acceptance of erroneous words. Norwegian studies[2,3] apply trigrams (patterns of 3 letters) to detect non existing combinations (see fig. 1). Comparing their tool (called SCARRIE) and Word98 it was found that their tool reduces the number of words not recognized (bad flags, words not recognized by Word98 123 errors, SCARRIE 60 errors). On the other hand both programs do not differ in terms of errors spotted(*Scarrie*: 90, *Word98*: 92).

Fig. 1: Marking of Errors: the light bars represent the SCARRIE results, the dark bars the Word98 results. Bad flags are words not recognized.

The main cause of not recognizing a proportion of the words can be assigned to the lexicon's being small, having too few varieties in Norwegian compounds. The lexicon of the Norwegian study consisted of 360,933 word forms, organized in 72.626 lemmas. The number of words not recognized will decrease with increasing size of dictionaries. *TALO's Norwegian lexicon has grown into 970.000 word forms and therefore more correct words will be labelled as correct and the speller will stop less frequently. Note that these cases are fundamentally different from "errors missed".

"Errors missed" are erroneous judgements of either the trigram method or the permutation method. An artificial process tries to analyse unknown words and decides "there are no unknown sections in the word and therefore, it might be a

correct word". However the Norwegian study indicates that these decisions frequently are incorrect. The question we might ask here is: how should we proceed in spelling if we reject such an unreliable artificial analysis.

WHAT ABOUT ALTERNATIVES?

Erroneous words should be matched with a series of alternatives. These alternatives should be similar in form. But are they similar?

The English word "abandon" should be one of the alternatives for the error "abbandon". Other words are very similar too ("Abaddon"). Words are look-alikes, but also soundalike. In English words that begin with **p** or **t** the **p** is not pronounced and an incorrect spelling like "tomaine" instead of "ptomaine" is likely to be made[4]
The similarity demand applies to German too. The German word *Aschermittwoch* and its genitive should be the only alternatives for the error *Aschemittwoch*.

The choice between alternatives is up to the user, but he should be shown only relevant alternatives.

A speller's lexicon should include a wide varity of information. This information should be accessible independent of whether the information is at the very beginning or at the end of the lexicon. It is the language model of TALO's spellers that extracts this information from the lexicon. This model is language dependent and, therefore, the German model is aimed towards German words only and the Dutch model is aimed towards Dutch words only.

Sometimes the difference between a correct word and the error is larger than normal: *registreerapparaat* versus *registratieapparaat* (an error). TALO's design is fitted to find these deviations. For Dutch soundalikes also exist (gogelaar instead of the correct word goochelaar).

In English many identically sounding words are easily resolved.

Spell-right
Spell-rite
Spell-write

Example 8

Except for school children aged around 8-11 years only the last example does

have meaningful connotations[6]. Adult spelling would be more or less determined by the selection of connotations.

Given the variety in language each language should be governed by a language model. The *TALO spellers benifit from the hyphenator's model. Moreover huge linguistic data bases are efficiently scanned during spelling.

THE TEST FOR ALTERNATIVES

A Dutch erroneous text was created with many intentionally incorrect compounds. The text included errors that occur in newspapers and usually are made by journalists due to time pressure. Our aim was to analyse the words that were not recognized as errors and to analyse the suggestions that were returned by the speller.
One of the spellers to be tested was based on the ispell paradigm, a tool built by programmers. The ispell speller design goes back to the early development in computer technology. Ispell cuts words rather in pieces.
As discussed above compounds are often created by permutation. But can we detect lemma's that are (randomly) inserted?

The error boeregemeenschap (farmer community) misses an n-binding and we would expect to see the alternative "boerengemeenschap". However, a series of alternatives were produced that did not have any similarity:

gemeenschapsgevoel
kerkgemeenschap
kloostergemeenschap
werkgemeenschap
dorpsgemeenschap
molazwangerschap
schaambeenboog
zwangerschapsdiabetes
mechanoreceptoren.

Example 9

Except for the element "gemeenschap" (community) the compounds do not show any relationship and some actually are very strange. The suggested words are lemmas added to "gemeenschaps", but the lemmas themselves are not related to the original word.
For the erroneous word "belastingsbeperkende" (an incorrect s-binding) the fol-

lowing alternatives were given.

belastingsbesparende
belastingsbesparend
belastingsvriendelijke
belastingsvriendelijk
belastingvriendelijke
ontsluitingstijdperk
augmentatieplastiek
verbrijzelingsfractuur
hulpademhalingsspieren
kraakbeenverbindingen

Example 10

Some of the alternatives are even incorrect! In Dutch "belasting" can have two connotations: tax and physical or psychological load. Only for load an s-binding is allowed. The first two alternatives concern tax issues and therefore the alternatives are spelled wrong. The friendliness (vriendelijk) of a load is quite useless, so it denotes "tax" and 3 and 4 are misspelled. The last cases again are unrelated and were presented after adding an additional medical lexicon which appeared to be counterproductive. The most serious case is formed by the erroneous alternatives.
More alternatives got a space, e.g., "dorps hoofd" instead of the word "dorpshoofd". The lemma "eeuwers" in twintigste-eeuwers became "eeuw ers".

For a demonstration text with 125 errors, for only 26 errors a proper alternative was given, 66 errors were not recognized at all (errors missed), and no proper alternative was provided for about 33 errors (see fig. 2). Apple's OS X TextEdit and Dutch speller in QuarkXPress 6.1 approved a large proportion of the errors too (accepted as having been spelled correctly), while errors spotted with a hit were faily low (see fig. 2).
One of the Finnish language technology companies focussed on morphological techniques to permute words. We understand why such a technique was useful for the Finnish language and the Finnish Microsoft's Word. But we do not understand why Microsoft took these principles and utilized them to the Germanic languages such as an analytic language like Dutch (using prepositions instead of a case system).
An article in "Onze Taal" confirmed the above problems[6]. The same demonstration text was used to analyse Word2002 Dutch. 48 of the 125 erroneous words were correctly detected, 43 were not recognized at all (missed errors) (see fig. 2).

*Fig. 2: Marking of Errors: the light bars represent the OpenOffice results, thereafter come Apple's OS X TextEdit, QX 6.1 Dutch speller, Word2002, and *TALO's speller in the last position of each group.* ***Errors spotted with a hit*** *increase with speller performance.* ***Errors missed*** *are words not recognized (accepted as having been spelled correctly). This index decreases with increased performance of the speller. Due to its design Speller XT 3.0, using the *TALO speller engine, does not miss errors.*

One of the main problems also recognized by the Dutch developers in MSWord is the approval of unusual incorrect compounds, slips of the pen, half corrections made by hand, etc. The answer to the article's[6] question "whether spelling is controlled" should be negative for the permutation method. The basis for the Dutch spelling corrector was the so called Green Booklet, it consists of basic forms only, and totals 125.000 entries. This is not sufficient for every type of text, but was said to be meant for normal texts.

But how common or widespread is a text? Again our aim is to demonstrate mechanisms and the risks they carry. An example of an approved error is "kamedel" for the correct word "kameel". The spelling tool accepts the error because the lemma "edel" exists and combines with the word "kam". However, the error "kamedel" is meaningless. The auther of Ref. 6 recognizes that the speller does not detect a rather large number of the possible combinations. The suggestion for *IJsselmeer* would be *IJskelder*, *Erasmus* is a *Grasmus*, and *allesdoeners* (Jack-of-all-trades) becomes *flesopeners* (bottle opener). These mismatches are the result of the small size of the lexicon. But is *IJskelder* similar to *IJsselmeer*? Should these cases betray the underlying mechanism to generate alternatives? *IJssel* is rather different from the Dutch word "*ijs*" (ice). It seems that the unknown word was split into "*IJs*|*sel*|*meer*" and "*sel*|*meer*" was replaced by the lemma "*kelder*" (cellar). The pair *Erasmus* and *Grasmus* shows that also initial lemmas are exchanged. To test the similarity we entered *Ijsselmeer* (Ij was

entered deliberately instead of IJ) in our speller to get similar alternatives, but with the standard accuracy only *IJsselmeer* was returned. We had to decrease the accuracy twofold to get the second alternative *IJsselwerf* (IJsselwharf) (2 letters different). So speller's alternatives could be more look-alike than the word *IJskelder*.

Morphological permutation is the main reason so many errors are not recognized. Morphological permutation also is the reason why alternatives do not look like to the error. The results of most tests were quite similar. Even trigram methods do not really improve performance. However, spelling strictly as in *TALO's speller avoids missed errors (see fig. 2).

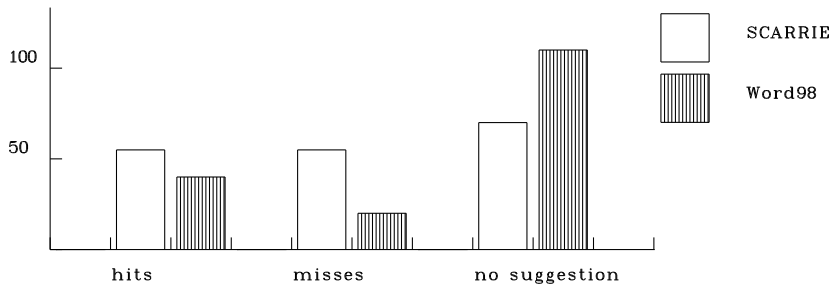

Fig. 3: Number of "hits", "misses" and "no suggestions" SCARRIE and Word98.

The results of the Norwegian study show that the main problem is the category "no suggestion" (see fig. 3). While Word98 scores fewer "hits" than SCARRIE it also shows fewer "misses". This is due to the relatively large number of "no suggestion".
No suggestions are justified if there is no relation between the error in the text with any of the words in the lexicon. However, by increasing the size of the idiom content it becomes more likely some match can be found, even for the more difficult cases such as *gogelaar* to be respelled as *goochelaar* when proper technologies are used. The column *no suggestion* in fig. 3 would be inversely related to the lexicon's size in terms of words. For the Norwegian language the SCARRIE example included 360,933 words, the number of words for the Word98 example has not been mentioned. *TALO's lexicon for the Norwegian language includes 980.000 words. Consequently *TALO's spelling engine can be expected to show fewer "no suggestions".

MULTIPLE WORD SEQUENCES

Some errors are generated through the effect of neighbouring words: *an house, an wive* (incorrect article), any human kinds (incorrect plural).
Sentences can be grammatically incorrect as in the Dutch example: "hun zijn de daders" ("Them are the culprits"), or "ik heb dezelfde hobby's als jouw."[7] ("I have the same hobbies as your"). This type of error may be caused by lacking formal language skills. Other cases are caused by a slip of the keyboard/mouse during the use of text processing tools. It is known that automatic spelling correctors even introduce error texts like: "De co-piloot ondernam niets ondernomen om de duikvlucht te onderbreken." ("The co-pilot undertook nothing undertaken to break off the nose dive.") The past particle explains that the word "heeft" ("has") was replaced by "ondernam" ("undertook"), but the previous past particle was not removed[8]. The number of possible errors is nearly unlimited, but some of these errors could be detected when the word boundary is crossed!
However, multiple word sequences which are used for comparisons should be considered carefully.
Language has its nuances. In English the position of "only" changes meaning:
Only she tasted the rutabaga (no one else did).
She tasted only the rutabaga (she tasted nothing else).
She only tasted the rutabaga (merely nibbled)[9].
Probably any grammatic analyser will fail to differentiate between these subtle differences.

In German the new orthography calls for multiple words comparison: 1) "Angst und Bange machen", against 2) "mir wird angst und bange". In the first case Angst and Bange have become nouns and therefore are written as majuscules, in the second case these words are adjectives and therefore written as minuscules.

CHANGE IN LANGUAGE

The Runeproject and Gutenberg text collection consists of author-right-free texts. These texts are at least 75 years old. There is also an author-right-free copy of the Webster's Unabridged Dictionary (1913 edition). This English dictionary also is over 75 years old. TALO language book[1] shows an example text from the Swedish Runeberg project and marks the words that have changed in spelling over time.
The spelling of the Gutenberg text collection and the 75 year old Webster also

do not agree any more with the spelling of the 3rd International edition of the Webster. The *Svenska Akademiens orflista över svenska språket* gives a reason for the changes. They introduced 5000 new words in the latest edition. They removed words such as *landskanslisten* and *militieombudsmannen* because the concepts or jobs do not exist any more and people do not drink *kaffesurr*[10] anymore. An other reason for changes in spelling are the spelling reforms. In 1990 a new recommendation was issued by the French Academie. The Dutch language was reformed in 1954 and again in 1996, and the German language in 1996. Many other languages gradually change their orthography over the years. Nevertheless outdated lexicons or lexicons based on outdated texts are still being distributed as parts of speller tools.

CONCLUSION

The performance of spelling is related to technology and to the lexical work required to cover a very large portion of the idiom of a language. This coverage should be accurate and not be based on a wild mix of non existing comparisons.

Spelling is not limited to isolated words. Words are used in context and context can be variable. As society changes, language changes too. Spelling tools should follow these developments in society and language (i.e. outdated lexicons should not be used). Spellers are meant for people who are uncertain about language. These people trust spellers, and it is up to the developers to add new linguistic technology to make users confident. Spellers should prevent the user from jumping from error to error, as is so often the case in daily newspapers.

The *TALO speller engines incorporate technology enhancements and the lexicons are kept up to date with changes in the respective languages. Keeping abreast of these changes and by continually meeting with people who speak these languages *TALO ensures that its products perform as expected.

REFERENCES

1 J.C.Woestenburg, *TALŌ's Language Technology, Hyphenation, Spell checkers, dictionaries, 2002, sec. edition, *TALO BV, Bussum (http:/-www.talo.nl/download).

2 Victoria Rosén og Koenraad de Smedt, SCARRIE: Automatisk korrekturlesning for skandinaviske språk, MONS 8, Tromsø, 1998.

3 Victoria Rosén og Koenraad de Smedt, Er korrekturlesningsevnen di god?, Resultater fra SCARRIE, Universitetet i Bergen og HIT-senteret, MONS 8, Tromsø, 1999.

4 J. Krevisky & J.L. Linfield, The Randon House Bad Speller's Dictionary, Random House, New York, 1991.

5 C. Croft, Write to Spell in Primary Classrooms, New Zealand Council for Educational Research, 1998

6 T. van den Heuvel, De Spelling onder controle?, Onze Taal, 2003, 9, 236-238.

7 J. van de Gein, "En toen kwam er een dief .....", Onze Taal, 2003, 11, 292-293.

8 R.Reinsma, Je zette zich schrap", Verschrijvingen op de computer, Onze Taal, 2003, 11, 312.

9 A.M.Siegal & W.G.Connolly, The New York Times Manual of Style and Usage, Radom House, New York, 1999, Back Cover.

10 Svenska Akademiens ordlista över svenska språket, Svenska Akademien, Norstedts, xiii-xiv, 2000.

Glossary

accuracy, the quality or state of being correct or precise.

compounds A concept consisting of more than one lemma, e.g., housemaster. In English a compound can be open, closed or the two lemmas can be connected with a hyphen, e.g., well-dressed.

compression, the reduction in volume by recoding redundant elements.

conjugated, conjugations, the formation or existence of a link or connection between nouns.

idiom the base of words having a distinct meaning in the language of a writer or any person or group that uses the language.

inflected, inflections, a change in the form of a word (typically the verb ending) to express a grammatical function such as mood, person, number, case, and gender.

lemma, the smallest section of a meaningful word, also root, stem, or lexicon entry.

linguistics, the scientific study of language and its structural components, such as grammar, syntax, and phonetics.

morphology, morphological, the study of the forms of words, in particular inflected forms.

orthography, the conventional spelling system of a language, how letters combine, represent sounds and form words.

permute, permutations, (verb) submit to a process of alteration, rearrangement, permutation; (noun) a way of getting all possible variations, in which a set of things (in our case words) is ordered or arranged. The number of permutations is the faculty of n (i.e. 1 x 2 x 3 x ..... x n).

syllable, a unit of pronunciation having a vowel sound with or without surrounding consonants.

trigrams, combinations of 3 succeeding letters that may occur in a target languages.

unfolded lexicon, a lexicon in which all encoded compacted cases are unpacked into their full form.