

# Hyphenation and spellchecking in InDesign, Smart Hyphen & Smart Speller

WoodWing Publishing Conference,  
Cancun, Mexico, November, 9 - 10, 2006.



*Jaap Woestenburg, PhD,  
\*TALÖ b.v.,  
Lijsterlaan 379,  
1403 AZ Bussum, NL.*

**\*TALO**

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

Copyright © \*TALŌ b.v., 2006.

All rights reserved. Without limiting the rights under copyright reserved above, no part of this production may be reproduced, stored in or introduced into a retrieval system or transmitted, in any form or by any means (electronic, mechanical, photocopying, recording or otherwise), without the prior written permission of both the copyright owner and the above publisher of this book

The greatest care has been taken in compiling this book. However, no responsibility can be accepted by the publisher or author for the accuracy of the information presented.

## 1. Introduction

Good morning Ladies and Gentlemen. I would like to draw your attention to the subject of language in publication, especially on tools that are said to do something with language. These tools are for hyphenation and spelling. These subjects also divide my presentation into its main sections. During this conference you must have noticed that WoodWing has released two new tools: Smart Hyphen and Smart Speller. To release a new hyphenator and speller is not without reason, therefore, the focus of this presentation is the performance of hyphenation and spelling, especially on the drawbacks of mechanisms and what the performance ought to be. In other words we will draw your attention to hyphenation and spelling errors! At the same time the presentation explains "*Why you do need a Hyphenator and Speller Plug-in for InDesign to Replace InDesign's hyphenation and spellchecker functionality*".

### 1.1. HYPHENATION

There exists the expression "a Hyphenated American". This is somebody who can trace his ancestry to another, specific origin in the world. However, if we look at printed hyphenation most people can NOT be traced back to their origin. "*Could hyphenation be made more reliable?*"

Let's begin with the hyphenation errors and where hyphenation fails. However, before going into details I have to state that most programs do use insufficient hyphenators, and the failures described below also apply to other text processing utilities being marketed.

A few examples of the hyphenator in InDesign and what it should be:

Standard InDesign	Smart Hyphen	key word to remember
<i>English:</i>		
dal~la•s•ite	dal~las~ite	†°
poly~o•n•y~mous	poly~on~y~mous	†°
polype•an	pol~yp~ean	† <sup>1</sup> , ‡°
<i>German:</i>		
in~tra•o•pe~ra~tiven	int~ra~ope~ra~ti~ven (new)	† <sup>3</sup> , † <sup>1</sup>
Pro~duktra~ti~o~na~li~sie~rung	Pro~dukt~ra~tio~na~li~sie~rung	† <sup>2</sup>
Ver~mö~gensauf~tei~lung	Ver~mö~gens~auf~tei~lung	† <sup>2</sup>
Wet~te•rin~for~ma~ti~onen	Wet~ter~in~for~ma~tio~nen	†°, † <sup>1</sup>
in~for~mier•t•heit	in~for~miert~heit	†°
<i>Danish:</i>		
re~sul~ta•t•in~di~ka~to~rer~ne	re~sul~tat~in~di~ka~to~rer~ne	†°
li~cen•s•om~rå~der	li~cens~om~rå~der	†°
na~tura•re~a~ler	na~tur~are~a~ler	† <sup>2</sup>
fa•ga~stro~nom	fag~astro~nom	† <sup>3</sup> , ‡°

In order of ranking (more seriously first):

‡°error, †°unstable, †² not hyphenated at compound boundary, †¹ not hyphenated, †³ widow

The unstable cases have 50% probability to be hyphenated erroneously. Why is hyphenation unstable? Why especially at the compound boundary? Why is the hyphenation performance so bad for many languages?

The answer is *the hyphenation technology*. This hyphenation technology goes back to the 80s and are based on Franklin Liang's dissertation on hyphenation (1983), sadly, the technology's assumptions are not "true". This hyphenation model accepts a certain probability of errors and it assumes that these errors are distributed evenly over the positions in a word, in other words, hyphenation is rather orderly (regular).

An example of regularity (Spanish) but don't miss the prefixes and their irregularity in Spanish:

com~pa~gi~nán~do  
 com~pa~gi~nán~do~la  
 com~pa~gi~nán~do~le  
 com~pa~gi~nán~do~lo

However, language does not behave regularly! Irregularity in Spanish:

Spanish su~b.. or sub~..., super~... or supe~r....  
 su•bas~ta~re~mos  
 sub•a~rren~da~ba  
 su•bas~ta  
 su~pe•ra~bles  
 su~per•a~bun~da  
 in~ter•ac~ción                      InDesign: in~te~rac~ción

Irregularity in English:

bib~li~o•phage                      [left or right of the ph]  
 bib~li~oph•ag~ic  
 bib~li~o•pole                      [left or right of the p]  
 bib~li~op•o~lism  
 bi~og•ra•pher                      [left or right of the g]  
 bi~o•graph•ic

Irregularity in German:

Ab~wasch•lap~pen [compound] left or right of the sch  
 ab~zu•schlach~ten  
 See for errors above.

## Irregularity in Swedish

kar~dio•skle~ros [compound] left or right of the skl InDesign: kar~di~os~kle~ros  
mänsk•lig

If we take a look at InDesign's standard hyphenator, it is unstable or erroneous at the compound boundary. The categories "not hyphenated", "unstable", "wrong" are all related to the uncertainty of InDesign's hyphenator. If uncertainty is too high hyphenation is suppressed, slightly less uncertain results in unstable or incorrect hyphenations. If developers of this hyphenator had known anything about language they should have had knowledge about the larger variability at the compound boundary. This means proportion of errors is NOT evenly distributed.

Whatever effort is made the standard hyphenator in InDesign cannot skip this erroneous assumption and therefore they are doomed to fail.

The TALO hyphenator (Smart Hyphen) does not assume an even distribution of errors or just accepts any proportion of errors. The TALO hyphenator is based on a language model which handles the irregularity at the compound boundary and elsewhere. This model is specific for any language and because the nature of language varies from language to language, the model varies from language to language too.

TALO hyphenator has been successfully applied in newspapers, ranging from Sweden, Iceland, Belgium, Indonesia, Africa, and many other countries. But being partner with Plug-In and XTension developers we run against the limits of e.g. InDesign, as in case of the Thai language, floating diacritics are rendered incorrectly, even when we use Windows XP's Thai modules. So language handling does not end with Unicode, but calls for proper positioning of characters too.

## 1.2. SPELL CHECKING

Orthography is the art of writing words with the proper letters according to standard usage; the correct spelling opposed to *cacography*, bad handwriting.

Let's continue with the second subject "Spell checking" and with orthographic errors NOT noticed by the standard spellchecker in InDesign. Again this spellchecker is not the only program which suffers from poor spelling algorithms.

Why has this algorithm never been replaced? We don't know, but most American only shop at home.

Recently *Bastian Sick of Der Spiegel* did put forward the question why people (in Germany) make so many errors with letters-in-between compounds (German, Fugenzeichen). A number of these compounds get an -n- in-between, others not. There are rules, but people have difficulty in applying these rules. You normally say "let's use the spellchecker". Yes, we use InDesign's spellchecker "German reformed".

Wrong → Correct  
 Bienestich → Bienen**st**ich  
 Fuge-Zeichen → Fugen**z**eichen  
 Oberklassewagen → Oberklassen**w**agen  
 Klasesprecher → Klassen**s**precher  
 Medikamentehersteller → Medikamenten**h**ersteller  
 Aids-Medikamente-Hersteller → Aids-Medikamenten**h**ersteller  
 halb-herzigen → halb**h**erzigen  
 wund-ähnliche → wund**ä**hnliche  
 Hirtestab → Hirten**n**stab  
 Klasse-Bücher → Klassen**n**bücher  
 Oberstufe-Schüler → Oberstufen**s**chüler  
 Pause-Brote → Pausen**n**brote  
 Breite-Sport → Breiten**s**port  
 Textiltapete-Kleister → Textiltapeten**n**kleister  
 Speisen**k**ammern → Speise**k**ammern  
 Instrumentekoffer → Instrumenten**n**koffer

According to InDesign's standard spellchecker all erroneous words are correct!, The following words were recognized as unknown: Sick (the proper name of author), Werbung!"-Aufklebers (why "Keine Werbung!"-Aufklebers), Duden (the proper name of a dictionary publisher), classis, E-Mails (correct), that's all. So InDesign does NOT spell check this text even if they say they do! The technology of InDesign's spellchecker starts with the assumption that compounds can be generated from single root words. In mathematics the results of such a process are called "Permutations", combinations of root words. Roots as Biene, Bienen (plur.) and Stich (der/-e(s),-e) exist. According to InDesign's assumptions

Biene•stich  
 Bienen•stich

But the first permutation is wrong, but approved by InDesign's standard spellchecker. The same technology is used for such.

amazonen•mier	amazone•mier
brief•bus	briefen•bus

All four words are assigned as correct by InDesign, but the first column is wrong! It is determined by meaning The *Amazones* were fighter women, the *Amazone* is a river in Brazil. *Briefen* is the plural of *brief* just like the German example. There is nothing wrong with InDesign's dictionary Dutch. It came from Van Dale (1995). It is a high quality lexicon, but the spellchecker algorithm downgrades the Van Dale lexicon, just as is the case for German.

These critical notes do apply to nearly all languages supported by InDesign. The

same critical notes can be applied to InDesign's competitors.

Permutations have another drawback. With increasing number of root words the total number of permutations increases sharply, and most calculators will produce a message "ERROR", the number overflows the digit's size. Therefore, constraint by constraint have to be set up to keep the spellchecker's process controlled.

The \*TALO spellchecker does not use permutations to create its lexicon. Instead it uses a very large lexicon which includes a large proportion of current day compounds. Together with a fast search engine the lexicon is scanned within a fraction of a second. Some lexicons have grown over the years to over 3 million words (compounds). Yet there still are unknown words, but keep in mind: *the user's idiom is also restricted*. The highly educated have up to 10,000 words in their active vocabulary. Women a bit more than men. A small section of these words are not in our lexicon. However an automatic learning mechanism based on history is supplied literally to catch these entries, and if necessary it is able to forget to prevent overflow. And we humans are easily overflowed. We can remember up to 7 items only.

## 2. Wrong Assumptions

We now return to the wrong assumptions and their consequences, however, any of the our statements also apply to a vast range of other commercial hyphenators and spellcheckers.

### 2.1. Hyphenation:

As we have seen the source of most of InDesign's hyphenation errors is the compound boundary. This is caused by a wrong expectation — the fluctuation of errors at the compound boundary agrees with fluctuation at other positions in the word<sup>†</sup>. I stress "this assumption is not true", but can it be solved by feeding information continuously. NO, the wrong assumptions violate languages time after time. Some words might be hyphenated correctly, other words, which were hyphenated correctly now are hyphenated wrongly.

---

<sup>†</sup> The Liang method as used by InDesign and others assumes an even distribution of hyphens and hyphenation errors over the positions in a word. In statistical terms this is named *linearity*. Any set of linguistic patterns is applied to all positions in a word. The patterns themselves are subject to uncertainty. Some patterns are certain other uncertain. Uncertain patterns are likely to be inaccurate and cause errors. Therefore a threshold is built in to prevent some hyphenation locations, however, the decision is still based on an uncertainty ratio and therefore subject to error. The amount of errors increases at the compound boundary. The incorrect assumptions result in a lot of errors while comparing apples and oranges.

## 2.2. Spell checking:

As we have seen the source of most of the orthographic mismatches is a very loose permutation technique. A lot of effort is spent to define constraints in order to reduce the infinity of combinations known from mathematics (calculate  $x!$  ( $x$  faculty) for  $x=1000$  on your calculator). Permutations create an extremely large lexicon. Some of the permutations are correct, but most of them just don't exist. A lexicon which consists of three-quarter of errors does not recognize three-quarter 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. You need to have a very large proportion of the full-size lexicon in order to differentiate between the wrong and correct.

## 3. A different approach,

In cooperation with WoodWing we have introduced Smart Hyphen and Smart Speller. The hyphenators accurately hyphenate compounds. Each hyphenator uses a language model which helps in determining the compound boundary. These hyphenators insert every possible hyphen except for widows. With a higher density of hyphenation it is easier to adjust the text over narrow columns. This higher density does not suffer from InDesign's/Proximity's uncertainty model. Deep inside, the hyphenator engines are compact and due to the language model a single processing path is necessary instead of determining the optimum of several possibilities. All hyphenators are stored in a single library, including the linguistic data.

The TALO spellcheckers use a strict technology to detect all vulnerabilities of a text and suggest alternatives that exist in language. We never say this combination of words might look OK, such a combination has to be notified to be OK. The lexicons cover the major part of the language idiom, fully conjugated or declined. The idiom is up-to-date! The spellchecker usually detects errors which were unnoticed (overlooked) by other spellcheckers (using permutation methods).

In respect of our technology we stress: "it is not engineering that matters in language processing but it is linguistics itself that determines how a process has to be organized. Hyphenators and spellcheckers can be made intelligently by setting up a human-like decision process. Humans do not hyphenate wrongly, computers do. Humans learn, but computer are caught in their own trap (permutations) time after time.

Is the base of traditional spellcheckers sufficient? No, a large amount of linguistic information is available from printed dictionaries, however, the inclusion criterion is "the usage over several years". This inclusion criterion has a delay, so many words that are used today are not yet found in the printed dictionaries, yet, a decision on the orthography has to be made. \*TALO maintains large lexicons and keeps them up-to-date. Since permutations do not help in finding errors, new technologies have been devel-

oped to handle large amount of linguistic data. Most Spanish and Italian verbs have 60 forms, while in other languages verbs take a few forms only. In Finnish, Estonian, and Hungarian prepositions are attached to the end of the nouns and adjectives, just like a case system, there are 13 or more cases instead of the 4 case system of German. Other languages hardly use cases (English only the possessive), the size of these lexicon is smaller but word order and collocations become important. Linguistically the differences between languages are large and the spellchecker has to take the behaviour of a language into account, so each language has to have its own descriptor instead of blindly using computational power. An *amazonemier* (amazon ant) is not a female fighter creature.

#### 4. Troy and the Amazon(s), the epilogue of the word's origin

The Amazons (fighter women) fought on the side of Troy in the Trojan War, and what did men do: Achilles killed Pen•the•si•lea, the Amazon's leader, Theseus won the Amazon Antiope (or Hippolyta), and their son was Hip•pol•y•tus<sup>†</sup>. The word Amazon means a- (not) + μαζός (breast), having only one breast, an advantage to shoot arrows during horse riding.



Misunderstanding the word's origin leads to poor shots<sup>‡</sup>.

In German the word "*Amazonenspringen*" is side saddle jumping on a horse. It is not

<sup>†</sup> The standard hyphenator of InDesign misses syllables in Pen•the•sileia (not a silent) and Hip•poly•tus (poly is not a prefix)

<sup>‡</sup> In Brazilian Portuguese the amazon woman is an *amazona*. An *amazonense* is a native inhabitant of the Brazilian state *Amazonas* and *amazonita* is a green mineral (feldspar). However the river (rio) is also named *Amazonas*. Except for Norwegian the river has been named the *Amazon*. In English *Amazonas* is a Brazilian state. However, an *Amazonasdelta* does not exist because the state of Amazonas does not touch the ocean. These fine differences have not always been kept intact in all languages or are used incorrectly.

"*Amazonespringen*" (as InDesign believes, accepts). The river is named the Amazonas. An error as "*Amazonengebiet*" is easily accepted, but a spellchecker should know the difference between the women and the river.

In English an Amazon ant is not a women to share a bed of flowers, unless you are Achilles.

In French *des Amazones fa•bu•leu•ses, ne sont pas (are not) la Région géo•gra•phi•que de Amazonie* (In French silents (•) are preferably not hyphenated, one of the French peculiarities).

In Norwegian "*det Amasonefolk*" is a "*sterk og stridlynt kvinnefolket*" but not related to the big "*Amazonaselven*" (river). InDesign's answer "Am\_ as\_ onasfolket" must be considered as complete rubbish (the underline symbols a space) and "*Amazonasfolket*" is also wrong, but InDesign's permutation mechanism wrongly accepts this erroneous word.

In Swedish the river is the *amasonfloden*, but InDesign also wrongly accepts *amasonenfloden* as correct and the ruler (drottning) of the Amasons is incorrectly hyphenated as *ama~~sond~~rott~~ning*, and the error just is at the compound boundary.

In Finnish the river is named *Amazonjoki*, the Amazons are *amatsonit*, the ants are named after the mythical ant warriors "amatsonimuurahainen" (Achilles' Myrmidons) and not after the river, unknown to InDesign's spellchecker.

In Czech these words do not compound and just upper case or lower case makes the difference Amazonka (the river), amazonka (one of the fighter women).

The form of compounds depend on meaning of the words and does not behave according to any permutational rule. Ignoring these facts makes spellchecking a farce.

Language did change with the passing of the years and it continues to do so. Therefore, the libraries and the lexicons you will use, are kept up-to-date by \*TALŌ. Knowledge about languages is fundamental in keeping them up-to-date.

## 5. Translation of words and terms

### English

dallasite,	a native resident of Dallas
polyonymous,	having many (poly) names (onymous)
polypean,	related to or like a polyp
bibliophage,	bookworm
bibliopole,	a dealer in books
bibliographer,	one who writes about books, their authorship, publications and similar details

### German

intraoperativen,	internally (intra) working (operativen dative or accusative case)
Produktionalisierung,	The Rationalization of a Product
Vermögensaufteilung,	distribution of wealth
Wetterinformationen	information (plural) on the whether
informiertheit	being informed
Abwaschlappen,	discloths
abzuschlachten,	to slaughter
Fugenzeigen,	characters in between used in compounds
Bienenstich,	bee bite
Oberklassenwagen,	top-class car
Klassensprecher,	class representative
Medikamentenhersteller,	producer of medicines
Aids-Medikamentenhersteller,	producer of AIDS medicines
halbherzigen,	literally half lovely
wundähnliche,	wound-like
Hirtenstab,	shepherd's crook
Klassenbücher,	class registers
Oberstufenschüler,	higher class student
Pausenbrote,	lunch/break (bread)
Breitensport,	popular sport
Speisekammern,	diner room
Textiltapetenkleister,	textile wallpaper paste
Instrumentenkoffer,	instrument case
Amazonenspringen,	ladies horse jumping tournament

### Danish

resultatindikatorerne,	indicators of result
licensområder,	license (licens) areas (områder)
naturarealer	nature areas
fagastronom	professional (fag) astronom

**Norwegian**

amasonefolket            the amazone women  
 Amazonaselven        Amazon river

**Swedish**

kardioskleros            cardiac sclerosis  
 mänsklig                humanlike  
 amasonenfloden        Amazon river  
 amasondrottning        amazon queen

**Finnish**

Amazonjoki              Amazon river  
 amatsonit                amazons (women)

**Dutch**

amazonemier,            amazon ant  
 brievenbus,              mailbox

**Spanish**

compaginando(la|le|lo),    from verb *compaginar* bring in accord  
 subastaremos            form verb *subastar* put up for auction  
 subarrendaba,            from verb *subarrendar*, sublease  
 superables,              (adj. pl.) surmountable  
 superabunda,            from verb *superabundar* frequently occurring  
 interacción,              interaction

**French**

Amazones fabuleuses,    the famous Amazons (the fighter women)  
 Région géographique de Amazonie    Geographical region of the Amazon (territory)

## Smart Language Tools

Smart Language Tools are Hyphenator and Speller Plug-ins for Adobe's InDesign meant to replace InDesign's standard hyphenation and spellchecker functionality. These language tools have been introduced to increase the performance of hyphenation and spell checking especially to avoid the drawbacks of InDesign's standard language support.

The standard hyphenator technology goes back to the 80s and is based on Franklin Liang's work (1983). This hyphenator model incorrectly assumes that errors and hyphens are evenly distributed over positions in a word, in other words, hyphenation is rather orderly (regular). However, errors and hyphens are NOT distributed evenly! The latter is related to language itself. This incorrect assumption is the main cause of wrong hyphenations, especially at the compound boundary.

The standard spellchecker accepts orthographic errors caused by the internal permutation algorithm. This algorithm is used to create artificial compounds on the fly, but the major part of these compounds does not exist. A serious drawback of this method is the acceptance of erroneous compounds which should evoke an alarm, but which are accepted as correct during spell checking.

For both hyphenation and spell checking a series of examples will be showed to point to the cause of the dysfunction. It is demonstrated that a different approach is needed to avoid the standard tools' incorrect assumptions: Smart Hyphen and Smart Speller use a language specific model to guide hyphenation and spell checking to a relevant accurate solution. Moreover spell checking does not rely on root words, but on a large lexicon covering the real language idiom.

In the final conclusion the word's origin is stressed. An example focussing on Troy and the Amazon(s) highlights the necessity for accurate linguistic information. The mythical fighter women (Amazons) who joined Troy are not related to the river Amazon, but .....

Title: Smart Language Tools ..... to bypass the drawbacks of InDesign's standard language support  
Jaap Woestenburg PhD,  
\*TALO bv.,  
Lijsterlaan 379,  
1403 AZ Bussum,  
The Netherlands  
<http://www.talo.nl/>  
see also: <http://www.woodwing.com/smartlanguagetools.htm>  
e-mail: [jaapw@talo.nl](mailto:jaapw@talo.nl)

## **A short biography of Dr. J.C.Woestenburg,**

Jaap Woestenburg graduated in Psychophysiology in 1978 at the University of Utrecht. In 1983 he received his PhD in social sciences on the subject of "Habituation of the Event Related Potential. A new technology.", on statistical models for the Neural Sciences. Thereafter he taught and conducted research at the University of Utrecht and later at the Free University in Amsterdam. During this time he studied the human performance and cognitive skills, especially related to the function of the brain. His work was published in international journals. Due to the explosive development of computer technology a contribution to technology and human-like processing was set-up. In 1985 he applied his knowledge about human mental process to the subject of hyphenation. The hyphenator tackled the major problem of many languages: the compound boundary and the structure of neologisms. The new technology was innovating and evoked the interests of WordPerfect. In 1986 the Dutch pattern based hyphenator was implemented and WordPerfect 4.0 primed the market with full-featured hyphenation which was unprecedented at that time. In 1987 \*TALO trademark was granted and in 1988 the company TALO bv was founded. After that a long period of research on many languages followed which formed the base of the current language hyphenation and spelling tools for more than 70 languages. In 1994 a cooperation with Swedish companies resulted in TALO's entrance in the pre-press industry. Later on a cooperation with Dutch and American companies resulted in a firm position on the spell checking and hyphenation market. In 2005 a request of an Icelandic newspaper called for the best Icelandic hyphenator for InDesign/InCopy. The Icelanders' request was the start of a cooperation between WoodWing and TALO. Since then many companies opted for the TALO solution. Next year our Enterprise will celebrate its 20 year anniversary.