*TALO 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

Newsletter

Married Words violently divorced
Hyphenation, what did you expect?

A high quality hyphenation function is vital for publishing, but isn't your hyphenation function promoted too optimistically? Things have got better, as time has passed. Yes, computers have become more powerful and performance has become better, but has hyphenation changed accordingly?
In earlier times computer dictionaries were small, and computer applications were rather restricted in usage. With so few words to hyphenate hyphenation patterns did not really compete/conflict with each other. Nowadays computers are used for nearly everything, language idiom has grown-up along with today's complexity of society. Competition between hyphenation patterns themselves now outnumbers the capabilities of most hyphenator's algorithms.

Maiden speech words, neologisms
The logical consequence of the above should be *mistrust*! Where would you expect a hyphen in a word? Given the widely spread usage of the Liang hyphenator model[1] a critical view of its performance is necessary.
Throughout the years we mustered large test corpora and many neologisms. These neologisms are essential because this linguistic data is statistically independent from any original data base used to develop a common hyphenator model. Neologisms supply us with a mean to estimate hyphenators' reliability. It also has to be said that building hyphenators with different data sets results in varying performances. Some do better, others do worse, but all are limited. Nevertheless the independent performance test does tell us something very important about hyphenation reality.

What can be said about performance?
The size of the pattern linguistic data base is a measurement of hyphenation performance (quality), the less the size of the data set, the worse the performance. If algorithms are based on incorrect assumptions, performance is always limited. If linguistic patterns heavily compete with each other, many sections within words will not be hyphenated. This applies almost certainty to English hyphenation.
Languages also vary in complexity. This complexity certainty has its drawbacks on hyphenation.

Language	PatternSizeTalo	PatternSizeLiang	Errors (in words) #%
NL	96183	116507	69,891 (of 742,700), 9.4%
DE	86374	50268	163,777 (of 972,127), 17.6%
UK/US	178005	117293	116,902 (of 318,307), 37%

Table 1.: Size of the hyphenation patterns and the number of erroneous hyphenation of a test corpus for Dutch (new), German (new) and English (UK/US). Non-hyphenated syllables are considered as errors too, as was abundantly clear in case of English hyphenations.

A Liang hyphenator model was used to hyphenate our Dutch, German and English dictionaries. Thereafter *TALO's test hyphenator compared these Liang hyphenated words with our own hyphenated corpora. None of these corpora have been used to develope Liang hyphenation patterns. Therefore these tests do not favour any comparison in advance.
Using these very large corpora, the number of hyphenation errors and omissions of the Liang algorithm proved to be substantial (see table 1.).

The erroneous hyphenations can be classified in three types of causes:
a) nearly all errors in Liang model occur at compound boundaries
b) there are three error classes:
- instability (2 hyphens around a compound boundary),

*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.

- mistake (an incorrect position of the inserted hyphen),
- omission of a hyphenation insert.

c) pattern size of Liang hyphenation tables considerably differs between languages (DE versus NL)

Examples of problematic **German** hyphenation:

Ab~fahrtss~pek~ta~kel	Ab~s~chöp~fungs~quo~te	Ein~sat~zein~heit
Ab~gren~zungs~wa~hn	Alu~mi~ni~u~mer~zeu~ger *	Fis~chauf~zucht~be~trieb
Ab~riss~s~topp	Ab~stieg~sängs~ten *	Fond~stöp~fe †
Ab~s~chied~s~tour	Ar~mee~e~in~sät~ze	Ge~richtsa~real
Ab~s~chlags~hö~he	Atho~sklos~ter	Nach~bar~p~latz
Ab~s~chluss~kos~ten	Bahn~hofsar~chi~tek~ten	Pro~t~es~tauf~kle~ber *
Ab~s~chluss~zah~len	Dre~her~laub~nis *	Ri~si~ko~bera~ter

Omissions can be vowel-vowel cases (hiatus), e.g., "stu-dien" instead of "stu-di-en", or can occur between consonants, e.g., "Wunschna-men", "Zen-tralla-bor" instead of "Wunsch-na-men", "Zen-tral-la-bor". An interesting feature is the large amount of errors in German, especially at the compound boundary. This large amount is related to the relative small size of the hyphenation patterns. Probably the German patterns have been calculated on a relatively small corpus.

Examples of problematic **Dutch** hyphenation:

af~slan~kope~ra~tie *	die~radop~tie *	kraams~ui~te
an~ti~kan~ke~rei~wit *	kee~t~jon~ge~re *	naakts~can~ner *
ar~moe~der~eis *	zon~ne~ce~l~in~du~strie	tui~ne~ve~ne~men~ten *
arts~enexa~men	wes~t-oost~ver~bin~ding	teflon~bal~le~tjes
bi~lim~plan~taat *	vluch~te~lin~g~en~quo~ta	taal~s~lij~ta~ge *
botoxin~jec~tie *	un~der~groundsce~ne	ja~rent~ach~tig~zan~gers *
bu~si~nes~sloun~ge *	kleitrek~ker *	fair~t~ra~de~keur~merk *

For Dutch the result was less extreme, but still 9.4 percent of the neologisms were incorrectly hyphenated (wrong position "bu~si~nes~sloun~ge" or not hyphenated "bo~toxin~jec~tie), most frequently seen in compounds.

Examples of problematic **US/UK English** hyphenation:

blowlamp	La~p~land *	cry~obi~o~log~i~cal
glareshield *	load~s~man *	looses~trife
in~fras~truc~ture	movieland	yup~pi~eness *
in~fundibu~lar *	yel~lowknife	scle~r~ob~last
in~trais~land *	Cey~lone~se	scrimshankers
in~tramem~bra~nous *	chronos~tratig~ra~phy	sul~fan~ti~mon~ic
in~tramer~cu~ri~al *	clado~ge~n~e~sis	sul~famet~hazine

For US/UK English more serious failures were observed. A third of the words in the test corpus were hyphenate differently, many hyphenations were erroneous, and an extreme proportion was not hyphenated at all ("movieland").

Thresholds keep errors from being seen

It might be possible that some incorrect hyphenations can be suppressed by increasing the hyphenation threshold, but very probably it will not bypass the real problem, as can be seen from hyphenation results in InDesign CS4 (*)[2,3]. Half of the erroneous hyphen locations concern compounds. Similar results can be expected from other hyphenators.

Performance factors

What is the effect of the pattern set and language models on hyphenation performance? In general TALO patterns are more compact than Liang patterns, the TALO hyphenator also produces very few mismatches on words presented for the very first time. Such a test is statistically independent of earlier calculations on known corpora. Usually error rates based on own corpora are presented, but these corpora do not predict hyphenation performance in respect of new words. The result of Liang patterns on independent new words (neologisms) is very poor. It is probably caused by incorrect assumptions — the underlaying linearity of the mod-

el. Moreover, the technology itself behaves like a picket fence of a very few pixels while viewing the Greater World[2]. Therefore, due to this mismatch, a lot of words are not hyphenated at all.

Language Model

Applying blind computational power doesn't result in better hyphenation. *TALO's hyphenator model is an accurate method to hyphenate words in accordance with national hyphenation rules[5]. Each language has its own hyphenator model and linguistic patterns are designed to detect compound boundaries. For English hyphenation, density is ca. 20% better than Liang hyphenation model and instability of patterns does not exist. In general a better density of text is also observed with other languages.

Summary

The Liang model is based on linearity and competing pattern. The principles of linearity do not match the way compounds are built up. Competing patterns based on a scale of a few steps are not very successful either. None of the disadvantages apply to *TALO's language models. The result is better hyphenation, both in regard to accuracy and density, i.e., less white rivers in text, less space needed to print the text.

References:

1 Liang. M., Word hy-phen-a-tion by Com-put-er, PhD thesis, Standford University, 1983.
2 Dr.J.C.Woestenburg, *TALO's LANGUAGE TECHNOLOGY, A Note on Hyphenation, 2005
3 Dr.J.C.Woestenburg, Hyphenation and spellchecking in InDesign, Smart Hyphen & Smart Speller, 2006
4 Dr.J.C.Woestenburg, *TALO's LANGUAGE TECHNOLOGY, HYPHENATORS, SPELL CHECKERS, DICTIONARIES, 2010
5 EuroAsiaHyphenatorUnicode 6.2.2, a Unicode hyphenator demo (multiple languages), see http://www.talo.nl/, menu download, section hyphenators.

Annex

Correct hyphenations

German:

Ab~fahrts~spek~ta~kel	Ab~schöp~fungs~quo~te	Ein~satz~ein~heit
Ab~gren~zungs~wahn	Alu~mi~ni~um~er~zeu~ger	Fisch~auf~zucht~be~trieb
Ab~riss~stopp	Ab~stiegs~ängs~ten	Fonds~töp~fe
Ab~schieds~tour	Ar~mee~ein~sät~ze	Ge~richts~are~al
Ab~schlags~hö~he	Athos~klos~ter	Nach~bar~platz
Ab~schluss~kos~ten	Bahn~hofs~ar~chi~tek~ten	Pro~test~auf~kle~ber
Ab~schluss~zah~len	Dreh~er~laub~nis	Ri~si~ko~be~ra~ter

Dutch:

af~slank~ope~ra~tie	dier~adop~tie	kraam~sui~te
an~ti~kan~ker~ei~wit	keet~jon~ge~re	naakt~scan~ner
ar~moe~de~reis	zon~ne~cel~in~du~strie	tuin~eve~ne~men~ten
art~sen~exa~men	west-oost~ver~bin~ding	tef~lon~bal~le~tjes
bil~im~plan~taat	vluch~te~lin~gen~quo~ta	taal~slij~ta~ge
bo~tox~in~jec~tie	un~der~ground~scene	ja~ren~tach~tig~zan~gers
bu~si~ness~loun~ge	klei~trek~ker	fair~trade~keur~merk

English:

blow~lamp	Lap~land	cry~o~bi~o~log~i~cal
glare~shield	loads~man	loose~strife
in~fra~struc~ture	mov~ie~land	yup~pie~ness
in~fun~dib~u~lar	yel~low~knife	scler~o~blast
in~tra~is~land	Cey~lon~ese	scrim~shank~ers
in~tra~mem~bra~nous	chrono~stra~tig~ra~phy	sulf~an~ti~mon~ic
in~tra~mer~cu~ri~al	cla~do~gen~e~sis	sul~fa~meth~a~zine

Author: Jaap Woestenburg, PhD, jaapw@talo.nl, Bussum, NL