*TALŌ's LANGUAGE TECHNOLOGY

A Note on Hyphenation

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, October 2005

ABSTRACT

Most hyphenation programs are based on a linearity principle.
According to this linearity principle the chance of an erroneous hyphenation is distributed evenly (= linearly) over the positions of letters in a word. In general, however, the syllables in a word are not distributed evenly over the positions of the letters in a word. In this Note, the linearity methods proposed by Liang and by other researchers are discussed. It is made clear why and how these methods lead to erroneous hyphenations, or, sometimes, incorrectly, to no hyphenation at all.
In the layout of a page these faulty hyphenation methods cause undesirable white rivers and irregular right margins.

SAMENVATTING

De meeste woordafbreekprogramma's zijn gebaseerd op lineairiteitsprincipe.
Volgens dit lineairiteitsprincipe is de kans op een foutieve woordafbreking gelijkmatig (=lineair) verdeeld over de plaatsen van de letters in een woord. De lettergrepen in een woord zijn echter in het algemeen juist niet gelijkmatig verdeeld over de plaatsen van de letters in een woord. In deze Notitie worden de lineairiteitsmethoden van Liang en van andere onderzoekers besproken en wordt duidelijk gemaakt waarom en hoe deze methoden tot foutieve, of soms zelfs tot helemaal geen, woordafbrekingen leiden.
In de opmaak van een pagina veroorzaken deze gebrekkige woordafbreeksmethoden ongewenste witruimten en onregelmatige rechterkantlijnen.

1. **Hyphenation**

"The hyphenation dictionary/algorithm proves to be efficient up to a certain degree (like in any program). Some words are properly hyphenated (or else, all suggested hyphenations are OK). Some others are partially OK. Some are wrong. This is not new and Scrubus ....." (Louis Desjardins, internet communication, 2004). In the additional information some basic errors in hyphenation for French are mentioned, but going after each and every word would be impracticable.

This type of behaviour applies to many hyphenators. Quark XPress 6.1 does not hyphenate French mutes (muettes) at least on the face of it. However this is not completely true. QX does not hyphenate the last 2 letters and therefore many muettes of two letters (at the end of the word) are prevented from hyphenation (bru~nette), but not muettes as "prac~tique" which QX hyphenates as "prac~ti~que". Some words with two letter endings are not muettes, e.g. "coupé", and should be hyphenated as "cou~pé", but Quark XPress 6.1 doesn't hyphenate this word.

Other hyphenators produce hyphens at both sides of a single consonant "ge-s-on-de" instead of "ge-son-de" (Afrikaans, Machteld Fick[1]). The model she uses is based on the mechanisms of human brain and is called a "neural network". Principles of these models might be valuable for visual perception and related pattern recognition, but the theoretical assumptions do not match to language. The microstructure of neurons does not necessarily behave identical to the macro level of language itself.
Hyphenation of words in texts has a long history.
Plain hyphenated-word lists have been used to hyphenate texts in documents (Troff on the Unices, PageMaker for the pre-press sector).
Liang[2] introduced a linear pattern recognition technique, which technique was applied to many languages (Boot[3], Tutelaers[4]). Knuth[5] claims that the Liang algorithm can be adapted for other languages by computing new patterns.
Daelemans[6] has shown that this is not true, because complete dictionaries of many languages are impossible to construct.
According to Nunn[7]) hyphenation programs based on pattern matching are essentially list-based as well. A draw-back is that patterns which are correct in derived words (the database of words which was used to calculate the patterns) but may produce incorrectly placed hyphens in words yet unknown to the system. In other words, this linear pattern recognition technique is not generally accurate. Considerable exception lists have been added to increase the accuracy. However, if pattern recognition incorrectly hyphenates in derived words the ex-

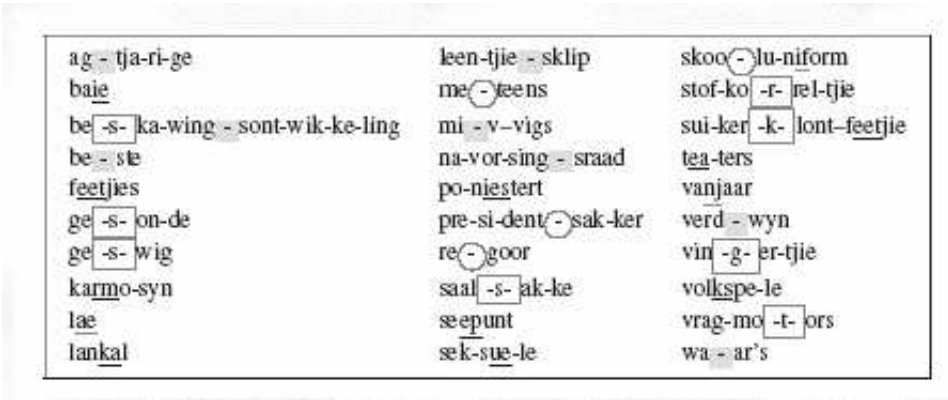

Fig. 1.: An Afrikaans Hyphenator doubles hyphens due to instability around a syllable boundary (marked as a rectangle, be-s-ka-ving.... instead of be-ska-ving....); and (marked with an underline) misses a hyphen, and inserts other hyphens incorrectly (marked with a circle).

ception list will continue to grow with time.
The method to compute Liang's pattern has been published in detail and because of its free availability the technology was widely applied in professional programs (QuarkXpress, InDesign, MSWord, etc.).

An aspect Liang and other authors didn't take into consideration are the assumptions underlaying the hyphenation method. Incorrect assumptions might violate the structure of the language and therefore lead to bad results. Therefore hyphenation pattern recognition should be based on algorithms that do not violate underlaying assumptions in language.

As will be demonstrated below, the structure of hyphenation of a language is too complex for most hyphenation techniques.

1.1. **The morphology of language**

A hyphen is a sign (-) used to join words as in "pick-me-up" and rock-forming". or to divide the syllables of a word at the end of a line.

The position of the hyphen mark varies and sometimes the variance of position is considerable.

bi.o.met.rics
bi.om.e.try
bi.o.log.ic.al
bi.ol.o.gist

Even in the English language hyphenation strategies differ between famous dictionary publishers. Hyphenation varies within a language[8,9,10]. These variations arise from the preferences of the different Style Manuals. In this article we will look at words with as many syllables as possible, regardless of Style Manuals.

The European languages can be grouped into clusters of languages, which are branches of the Indo-European family. Each of these branches has its own characteristics which determine hyphenation. Some languages do not belong to the Indo-European group but for centuries have been influenced by their neighbours.

Most of the European languages are strongly inflected. The case system adds endings (dative, accusative, locative, etc.) which in English would look like:

house.at
house.to
house.in
house.above
house.below

This type of ending occurs in Finnish and Estonian. These languages add prepositions to the end of the word and if youngsters (Fins and Estonians) have to learn a foreign language like English they have to memorize the concepts in rows like the above examples.

The Scandinavian languages add the article to the end of the word. For instance "hus" (which means house):

hus
hus.et
hus.er
hus.erna

It should be noted that these morphological endings do not necessarily determine hyphenation.

Many languages use compounded words. The mechanism to bind words is limit-

ed to implicit rules usually set by meaning. Compounded words can be long.

German:
Ab.fall.ent.sorgungs.satzung
Agrar.ab.satz.förderungs.durch.führungs.gesetz
Arznei.mittel.aus.gaben.be.grenzungs.gesetz
Donau.schiff.fahrts.gesellschaft

This variety of mechanisms in a language should be accommodated for in the mechanism of hyphenation.

1.2. **Linear methods**

In the 1980's Liang[2] developed an English hyphenator based on linear pattern comparisons. In Dutch the patterns look like:

1. lan4d3r
2. 4lann
3. l4do4p
4. er3t2h
5. 3t2hei

The patterns use numbers to represent a hyphen position. Odd numbers are acceptable, even numbers are unacceptable, higher numbers take precedence of lower numbers.
A few examples demonstrate the inadequacy of the generated patterns. The first of the five patterns listed above cannot choose between hyphen positions like: *eiland~raad , calan~drone, flan~dricisme*. The second pattern does not match with any word in *TALO's data base[11], and it would compete with words like *aanbouw~plannen*. The third pattern does not differentiate between words as *geld~opname, peul~doppen, meubel~dop* etc. The fourth pattern prefers *alert~heid, knoert~hard* over hyphenations like *kinder~theater*. The fifth pattern is more selective and catches all words with *...~heid*, but it leads to problems with words like *Elizabeth~eilanden*.
In most cases more than one candidate pattern needs to be considered in order to be able to decide which one to use. The set of patterns is linearly scanned, from one position to the next position, and so on, running through the word. In some cases the selection criterion leads to the correct solution, in other cases a wrong solution is chosen. As is illustrated above, the Liang model requires a competition between patterns but it is questionable whether competition can re-

solve every dilemma.
Statistical calculations were applied for generating the patterns, resulting in a set of probabilities. To deal with such a set of probabilities either the ranking or the averaging method can be used:

$$h_i = \sum p_i \cdot c_{i.j} \qquad i = 0, ..., n \ ; j = i, ... , i + m; p = 0, ... , 9^{\ddagger}$$

in which *i* is the linear index in the word and *j* is the linear index in the pattern, *c* a pattern itself, and *p* the probability of the pattern. The probability can only take a value in $0 \leq p \leq 9$ (from certainly 'no' up to certainly 'yes'). Its real value, as it would applied to the language itself, is truncated to a small integer number. Therefore the classification of hyphenation weights h_i used to hyphenate consists of a few 10% steps only. Taking the highest value in a winner-takes-all approach disregards any inhomogeneity in the words to be hyphenated.
The distribution of probability is the level of certainty (or uncertainty) expressed as a function of the position within words. At individual positions the uncertainty "whether or not to hyphenate" could be high. The model is based on probability which automatically implies the acceptance of errors. Usually those positions with a low uncertainty are accepted and the remaining positions are left unhyphenated.

There are two objections to this model:

a) the assumption of linearity
b) the implications of probability

ad a) The assumption of linearity implies that the probability of hyphenation is equally distributed over the word, which would be the case in (see fig. 2):

a.ba.ba.ba.ba.ba.ba.ba.ba.

This is very unlikely to occur.
ad b) The implications of probability tell us that one of the positions is more probable than the other positions, but this statement also tells us that there is the probability of being wrong. Once probability is accepted and categorisation only runs from 1,...,9 it will be impossible to get a finer degree of separation (less than 1 unit).

---

‡ This formula represents a moving average to emphasize the strongest hyphenation weights (*h*). *p* represents the unfiltered probability of the pattern *c*. To speed up the process one might choose a winner-takes-all approach, the pattern to be probably correct. Since the calculated patterns were based on probabilities errors are likely. The distribution of these errors of the position within the word is not affected by any selection criterion. The distribution of errors is related to the levels of hyphenation marks in the patterns themselves and to the sample size on which the patterns were calculated.

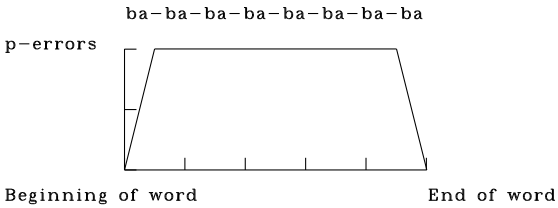

Fig. 2.: The distribution of errors according to the linear model. At the beginning and end of a word the distribution falls off to zero because hyphenation before the beginning and after the end are not possible. Within the remaining interval (horizontal section) the distribution of errors is equally high.

The integer steps to describe differences in probability lead to a considerable degree of uncertainty. They only differentiate in steps of 10. A considerable number of errors can be expected if all probability levels are used in a linear hyphenation model.

1.3. **The implications of incorrect assumptions**

If the linear assumption is not valid errors in hyphenation will be made. The invalidity of the Liang class of models can be demonstrated in German compounds as:

correct: .	incorrect:
Anhang.erläuterung	Anhan.gerläuterung
Dokumentations.abteilung	Dokumentation.sabteilung
Schwangerschafts.abbruch	Schwangerschaft.sabbruch

The incorrect cases follow a more regular distribution matching the distribution of a linear model. In fact quite often there exists a bimodal or trimodal distribution (compounds consisting of 3 or 4 root words). Accepting a linear distribution instead of a multimodal compound distribution increases the uncertainty at compound boundaries, which increases the variance at the compound boundary. Consequently, less reliable information is available to determine a correct hyphen location. The real effect is that uncertainty rises sharply where correct hyphenation is needed most (see fig. 3).
If the strategy is chosen "to suppress hyphenation when uncertainty is high" many compound boundaries will never be hyphenated. This is likely to result in white rivers running all over the printed page and/or the appearance of a highly irregular right margin.

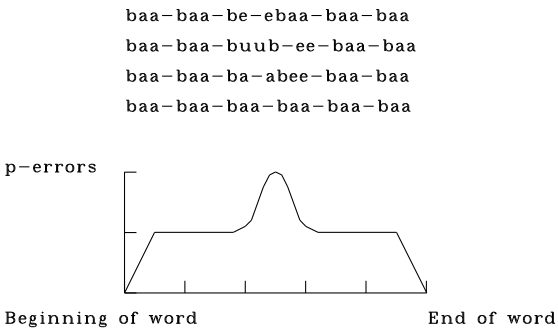

Fig. 3.: The probability of errors according to the linear model. The inaccuracy is enlarged around the compound boundaries. Therefore more errors are expected to be made at the bump of the curve than at other locations.

1.4. **Uncertainty and suppression of hyphenation.**

Nearly all implementations of Liang's model[2] use a probability value p to rank hyphen marks to differentiate between certain and uncertain cases. Since the length of an individual pattern is restricted, certainty becomes less than expected, which leads to to an unacceptably high number of hyphenation errors. In practice this inconvenience is resolved by "not hyphenating" above a particular uncertainty level. However, this method has its drawbacks. Due to the invalid assumption of linearity, uncertainty of hyphenation decisions rises at the compound boundary (see fig. 3). Consequently, compounds will be hyphenated in a lesser number of cases, resulting again in highly fluctuating white rivers on the page or in an irregular right hand margin.
Even a neural network like Fick's model[1] is linear and uses summation to stabilize the calculations, and therefore it is subject to the same type of errors.
A dramatic case of handling uncertainty during hyphenation is given in fig. 4. An open software package derived from Sun's StarOffice doesn't treat the text adequately. The text was printed in small columns.

Fig. 4.: Hyphenation in the Dutch language as suggested by an open software package. The hyphenator's result not only is erroneous, but, also, while keeping uncertain locations unhyphenated an awkward printing image is all we get. Words are divided unexpectedly and white rivers flow through columns. Each paragraph is defined by an indent and within a paragraph the lines should be filled up, but they aren't! (correct hyphenations: beademingsinrichtingen, beaujolaisstreek, bedrijfsinternetvoorzieningen)

2. **Conclusion**

Current linear hyphenation technology is largely based on the early work of Liang (1983), but results have been disappointing. This approach to hyphenation applies a standard mechanism to all kinds of languages, despite the very different morphologies of these languages.

One of the assumptions used by Liang and others, is linearity itself. Liang and others use a comparison method which goes step by step through a word. They assume that the probability of errors is evenly distributed as a function of position within a word. Given the structure of compounds in many languages it can be made clear that the syllable distribution and therefore the distribution of hyphens does not behave linearly. Consequently, this wrong assumption gives rise to errors at the compound boundary.

It could be demonstrated that accepting probability to calculate and to give weights to the pattern implies acceptance of hyphenation errors, or an increase in errors that is greater than it needs to be.

The decision not to hyphenate uncertain cases causes white rivers or irregular right margins and in case of narrow columns unacceptable conflicts which distort the lay-out.

REFERENCES

1 Neurale netwerke as moontlike woordafkappingstegniek vir Afrikaans, Machteld Fick, S.A. Tydskrif vir Natuurwetenskap en Tegnologie, 22, 1, 2003.

2 Liang. M., Word hy-phen-a-tion by Com-put-er, PhD thesis, Standford University, 1983.

3 Boot, M., Taal, tekst, computer, Servire, Katwijk, 1984.

4 Tutelaers, P., Herziene afbreekpatronen voor het Nederlands, MAPS 1993, pp. 187-190, 1993.

5 Knuth, D.E., TEXT and Metafont: New Directions in Typesetting, Bedford, MA: Digital Press, 1979.

6 Dealemans, W., AUTOMATIC HYPHENATION: Linguistics versus Engineering, IN: Worlds Behind Words, F.J.Heyvaert & F. Steurs eds., Leuven University Press, 1989.

7 Nunn, A., Automatic Hyphenation of Dutch Words based on Linguistic Rules, Van Dale Lexicografie, Utrecht, 1998.

8 Longman Dictionary of contenporary English, Longman, Harlow, UK, 1987.

9 The Oxford Colour Spelling Dictionary, Oxford University Press, New York, 1996.

10 Webster's Third New International Dictionary Unabridged, Merriam-Webster, Springfield, USA, 1993.

11 *TALO's Language Technology, Hyphenation, Spell checkers, Dictionaries, J.C.Woestenburg, *TALO bv, Bussum, 2002.

Glossary

Accuracy, the quality or state of being correct or precise, in hyphenation an inverse function related to hyphenation errors.

Assumption, statement that is accepted as true without proof.

Bi-, tri-modal, involving two or three modes, in particular having two or three maxima.

Compounds, a concept consisting of more than one root word, e.g., housemaster. In English a compound can be open, closed or the two root words can be connected with a hyphen, e.g., well-dressed.

Distribution, the way in which something is spread out among a group or spread over an area.

Hyphen, Hyphenation, a sign (-) used to join words to indicate that they have a combined meaning, or a division of a word at the end of a line. Placing hyphens is called hyphenation.

Indo-European, the family of languages spoken in most of Europe and Asia as far as northern India.

Linear, arranged along a straight, or nearly straight, line

Linguistic pattern, a language specific sequence of characters or letters found in written texts

mutes, muettes, a letter not pronounced, in French muettes as "brunette", pronounced as "brunet".

Neural Network, a computer system or program modelled on the human brain and nervous system.

Non-linear, not denoting, or not involving or not arranged in a straight line.

Pattern recognition, the detection of defined patterns to be used in a decision process.

Probability, the extent to which an event is probable; the likelihood that an event will happen. Probable events are certain, improbable events are uncertain.

Quarks, Quantum, in physics a subatomic particle, postulated as the building blocks for other particles. A quantum is a discrete, smallest amount of energy. These smallest particles are a metaphor for the smallest particles of meaning.