

2.2 Practical application to PCs

A personal computer has three main limitations when compared to a mainframe computer. These are its processing power, its memory size, and its disc capacity. These differences are decreasing, but that does not alter the significance of our approach. In addition, for the foreseeable future, personal computers will not be the final processors of text before quality publication, but only text entry and editing devices. Therefore the need for correct hyphenation is less stringent than on a mainframe computer.

The practical application of a hyphenation method can then be expected to be a program which uses a small amount of memory and is thus resident in the computer throughout the text processing operations. An exception dictionary should be optional, dependent on the disc storage available.

2.3 Practical application to computers

We assume a mainframe computer to be one which has large processing power, memory size, and large storage capacity. In addition, for the foreseeable future, these machines will be the final processor of text for quality publications with output to a phototypesetter, and will receive their information directly, or from other computers such as a personal computer.

The needs in this case are for correct hyphenation wherever possible, fast processing, and access to a dictionary in certain circumstances. The practical application can then be expected to be a program which uses appreciable memory and may be overlayed. A major dictionary should be optional.

2.4 Typographic considerations

In addition to the linguistic and computer considerations, there is also the question of typesetting to be taken into account. The whole purpose of the art of typography is to produce a visually attractive page where no element detracts from the readability of the text. Thus, over the years, printers have acquired rules-of-thumb that enable them

to avoid features that would otherwise mar the reader's fluent information gathering. However, it is important to remember that these are *conventions* only.

It is this typographic constraint that gives rise to the requirement that, in all orthographies using the roman script, no syllable may be visually divided.

In addition, printers specify the minimum number of letters that can appear on a line either before or after a hyphen, which is typically 2 on the first line, and 2 on the second. Thus A-POLITIEK, while correctly divided on a syllable boundary, is still unacceptable for our purposes. Examples of shortest breaks allowed are BY-GEDRA, SA-KE, AKADEMI-CI. Additional 4-letter examples from Die Bybel include: ge-sê (Matt 17:15), go-de (Joh 10:35), bo-me (Luk 21:29) and He-re (Joh 13:36). No words with fewer than 4 letters are broken in Afrikaans.

The word must also consist of two or more syllables, and so words such as SPREUK, STRING, and BLAAR may not be broken.

Because Afrikaans, like Russian and German, is an **agglutinative** language, we must allow for multiple compound words and 60 characters seems a reasonable maximum. For example

GE-MEEN-SKAPS-GE-SOND-HEIDS-VER-PLEEG-KUN-DI-GE

The handling of accents differs from language to language. For example, in Turkish both dotted and undotted forms of 'I' occur in upper and lower cases. In Afrikaans, accents occur in both cases in display material, e.g. advertising, but are generally avoided on capitals in normal continuous text. However, in high quality material such as Die Bybel, all accents are shown.

The following accents appear in Afrikaans

à á â ã è é ê ë ì í î ï ò ó ô õ ù ú û ý ÿ

À Á Â Ã È É Ê Ë Ì Í Î Ï Ò Ó Ô Õ Ù Ú Û Ý ÿ

Of these it should be noted that the accent ´ is used only for emphasis (dáár), in the same way that an italic typeface would be used. The accent ` has a conventional use also (òf ... òf). These two thus have no syllabic value and can be ignored for hyphenation purposes. However the diacritical marks " and ^ affect the phonetic values, and thus may be important for word division*.

We have been informed of a special accent usage in Afrikaans; when a word containing an accent is hyphenated and the accented letter is the first of the new line, the accent should be omitted†. This action can be implemented relatively easily by the text processing or composition program after applying the hyphenation algorithm. The program must just remove the accent or replace the character with its unaccented equivalent. However, when the text is subsequently modified and the line ending changes, the accent would have to be re-inserted automatically and we consider that this is quite difficult. Fortunately, this rule need not be addressed by us here, since it does not affect the hyphenation point.

In addition to these symbols, Afrikaans texts also contain various **ligatures**, which are one or more letters designed as a group. The group prints as a single character on phototypesetters, but most office printing devices do not have these features. Those combinations most commonly used in Afrikaans to provide ligatures are

ij 'n ff fi fl ffi ffl

The *ij* combination (originating from the Dutch) must be treated as the single sound [əi] in all contexts, but we have only found occurrences in proper names. The 'n only occurs on its own, and has no capital form.

All the others shown above are typographic features not affecting phonetic values. So that if the *ff* is internally represented by the ASCII character 249, this means that any hyphenation program will have to recreate the two original characters before tackling word division. Note also the obscure problem of hyphenating *ffi* into *f-fi* which thus generates a different ligature; fortunately the new ligature occurs after the hyphen and therefore appears on the second line where it does not affect the first line's inter-word spacing.

* See section 11.1 for later findings about the circumflex.

† Private correspondence from Nasionale Pers, 1987.

betreffende
sertifikaat
fluit
affiksie
souffleer

Examples of ligatures used in Afrikaans

The question of **natural** hyphens needs attention. These are hyphens which have been inserted by the author, and the rules for so doing are well agreed in Afrikaans (AW&S, 1964). The general rule, and the one we follow here, is that no word containing a natural hyphen may be divided again. This may produce very long unhyphenated words, and the rules can still be applied, but the composition program must then convert the word to one without a hyphen before presenting it for hyphenation.

A special case of the use of a natural hyphen exists in Afrikaans. When a disjunction (*afstandkompositum*) is formed, for example

HUIS-, SKOOL- OF WERKSITUASIE
or
KOSTEBEREKENING EN -BERAMING

then it is possible for hyphens to begin or end words. Here again we do not break these but suggest this is tackled manually according to house rules.

When it comes to actual insertion of the generated or **system** hyphen, then two forms are followed: the use of the short dash (-) and the use of the short equals (=). There was a movement towards the latter in Afrikaans for some decades (see figure below) but this is now reversing, and a summary of 26 Afrikaans books reviewed in the last year shows that 25 use the dash and 1 the equals sign for hyphenation. A quality indicator of the trend is Die Bybel which also uses the dash in the Nuwe Vertaling.

die oorspronklike draers van dié name al heel weinig gelykenis vertoon – tot ergernis van latere natuuronderzoekers soos Le Vaillant, Lichtenstein, Burchell en andere" (a.w., p. 9).

Van die diernaam is *wolf*, *tier*, *luiperd*, *jakkals* en *seekoei* besonder opvallende voorbeelde van misleidende naamgewing met betekenisverskuiwing. Ons noem in Afrikaans bv. die van die hiëna-soort *wolf*, terwyl die Afr. *wildehond* tot die hiëna-soort behoort. Die Afr. *tier* (ouer vorm *tiger*, vgl. *Tygerberg*) heet in Nederlands *panter* of *luipaard*, nie *tijger* nie; die dierscort wat in Ndl. *tijger* genoem word, kom in Suid-Afrika glad nie voor nie. Om die verwarring nog groter te maak, word in Afr. *luiperd* soms ook sinoniem met *tier* gebruik. Maar hierdie gebruik van *wolf* (= hiëna) en *tier* (= luiperd) is vandag in Afrikaans beperk tot sprekers van 'n ouer geslag. Die gangbare vorme is

= used as hyphen in Afrikaans text*

2.5 Why hyphenate at all ?

While there is a clearly identified need for proper Afrikaans hyphenation within professional typesetting organisations, in other environments the process of hyphenation may become redundant. Is hyphenation really necessary in Afrikaans ?

Various schools of typography approach hyphenation in different ways. There are those who follow William Morris and Eric Gill in believing that a "10-12 word line and even spacing between words are themselves of real & paramount importance, while the equality of length of lines is not of the same importance" (Gill, 1931). This may leave a ragged right-hand margin, which, depending on the width of the column and the type size being used, can result in minimal hyphenation being used. A widely-held view among typographers is that in fact text set in this manner should not contain hyphens at all, although Gill was very contented with them, as the following example shows.

the same whether for hand-cut punches or those cut by the mace. 're, tho' either of these processes can be done with or without machinery. Mechanical casting appliances offer a higher average of accuracy, and this is considered of paramount importance by some printers and publishers.
¶ Obviously the great if not the only advantage of mechanical punch-cutting is that once the pattern

The above extract was set with the same line breaks as used by Gill.

* E H Raidt, *Afrikaans en sy Europese verlede*, Nasou, 1980, p 103.

the same whether for hand-cut punches or those cut by the machine, tho' either of these processes can be done with or without machinery. Mechanical casting appliances offer a higher average of accuracy, and this is considered of paramount importance by some printers and publishers.

¶ Obviously the great if not the only advantage of mechanical punch-cutting is that once the pattern

The same material set fully justified.

An example of changing line width can be seen from the following examples re-typed from Beeld, 18 November 1987. On the left is the extract as it appeared in the newspaper. Three hyphens on succeeding lines are frowned upon in bookwork but this occurs frequently in newspapers, however, at the cost of extravagant spacing in lines 2 and 5. On the right, the same material is set in a column 5 points wider i.e. 3,5%, resulting in 2 fewer hyphens.

Die vrou in die koninkryk
deur Laura Steyn
(Lux Verbi, R8,95) is
nog 'n goed versorgde
slapbandbundel met
"boodskappe" wat vol-
gens die uitgewer spe-
siaal saamgestel is vir
gebruik by wykbyeen-
komste en vrouebid-
ure. Ek vind die bun-
del heel leesbaar.

Die vrou in die koninkryk
deur Laura Steyn (Lux
Verbi, R8,95) is nog 'n
goed versorgde slap-
bandbundel met "bood-
skappe" wat volgens die
uitgewer spesiaal saam-
gestel is vir gebruik by
wykbyeenkomste en
vrouebidure. Ek vind
die bundel heel leesbaar.

Some publications avoid hyphenation altogether, but this can produce grotesque results. In practice, text for reports, newspapers, books and magazines, requires most efficient use of the space on the page, and "provided that words are really distinct from one another, they should be set as close as possible" (Gill, 1931). However, this does not mean that justified type is preferred to ragged right setting. In this example from the same edition of Beeld we find a narrow column to fit round a headline. On the left is the text as it appeared in the newspaper with hyphenation inhibited, and on the right our re-typing with the hyphenation allowed. Beeld house rules dictate that proper names would not normally be hyphenated.

Super Fax en Altech se
Suid-
Afrikaanse
Ope sal
speel, het die
skeidsregter,
Richard
Kaufmann,
ná 'n
twyfelagtige
beslissing teen
die einde
van die tweede
stel in sy
kwarteindstryd
teen Anders
Jarryd van
Swede

Super Fax en Altech se
Suid-Afri-
kaanse Ope sal
speel, het die
skeidsregter,
Richard Kauf-
mann, ná 'n
twyfelagtige
beslissing teen
die einde van
die tweede stel
in sy kwart-
eindstryd teen
Anders Jarryd
van Swede

In a wider context, many people now come into contact with the problem of fitting text onto lines through the use of word processors. We investigated the approaches adopted by some of the major word processing packages and summarise our findings below.

- Displaywrite 3 Contains a hyphenation dictionary
- Easy Contains no hyphenation
- Macwrite Contains no hyphenation
- Multimate 3.3 Contains no hyphenation
- Volkswriter 2.2 Contains no hyphenation
- Word Contains optional logic for American and an exception dictionary
- Wordperfect 4.1 Contains no hyphenation
- Wordstar 2000+ Contains logic for American and semiautomatic hyphenation
- Xywrite III+ Contains logic for English and an exception dictionary

Generally, the occurrence of hyphenation is reduced as the line gets longer, as the type size decreases, or inter-word space gets smaller.

We can therefore conclude that for the professionals, hyphenation is essential; for the desk top publisher it is a 'nice to have' feature but not very important.

Chapter 3 PREVIOUS APPROACHES

3.1 Linguistic

Based on the agreed linguistic rules for *spelling* Afrikaans (AW&S, 1964), some computer implementors have attempted to codify orthographic rules from them. Hyphenation rules such as 'always break after a diphthong before a consonant followed by a vowel' can be quite useful, and are linguistically accurate. However, the semantic and phonetic implications of these rules imply considerable contextual analysis for the computer program, which is both difficult and unnecessary. We are, after all, only dealing with a text string.

Using the correct prefixes and suffixes such as those identified by Kotze (1979) and Marais (1970), and some of the compound boundaries also, the linguistic approach has been moderately successful when applied to computer hyphenation. However, this approach suffers from a lack of understanding of the typographic requirements. In addition, exceptions are difficult to identify and have not been exhaustively listed. Thus BE- is normally a prefix as in BE-DANK, but in BEDLAKEN it is not, nor in BEENVEL. To the mother-tongue speaker of course the solutions are obvious, but not to the computer programmer.

3.2 Algorithmic

We have identified 6 programs in South Africa which use an algorithmic approach. Often, a corpus of text has been analysed, usually inserting the hyphens manually first, and a set of rules, similar to the linguistic ones, has been propounded. In general our informants believe that no set of rules can possibly cater for all the constructs in the language, particularly as it is possible to create new words at will. In most cases, the logic of the method is unknown by programmers in South Africa or is inaccessible for proprietary reasons. Our findings are given below.

- A M International uses a Dutch hyphenation algorithm on all their systems.

- EAC Graphics (Compugraphic, CPT) uses an algorithm based on an early analysis of the language done by Linotype in the UK. The logic is not available in South Africa. It uses an exception dictionary built by the user.
- Monotype had an algorithm on their filmsetter system which was based on Dutch and German with corrections for Afrikaans. The logic cannot be easily deciphered and this system is no longer marketed.
- High-Tech Graphics (Scan Text) have an algorithm that was written in Germany but most customers are not happy with it and use a Dutch algorithm, storing exceptions in a dictionary.
- PhotraGraphics (Hastech, Linotype, Xyvision) use algorithms adapted from Dutch or Flemish versions, again with an exception dictionary. Interestingly, some recent work has been necessary for a new product which resulted in the re-drafting of the phonetic rules and sending these to an Afrikaans programmer working in the USA.
- Atex Systems' algorithm was based on a Dutch Algorithm, with some non-occurring clusters removed (Appleton, 1985; Forsyth, 1986). It is listed below.
 1. Since the word must contain at least two syllables, the algorithm searches backward from the end of the word for at least one consonant between two or more vowels.
 2. The word is scanned forward to find pre-defined common 'prefix' syllables. This routine is called recursively, until either no more prefixes occur or the end of the word is reached.
 3. The word is scanned backward to find pre-defined common 'suffix' syllables. This routine is called recursively if a suffix is found, until either no more suffixes occur or the beginning of the word is reached or the search reaches the end of a previously determined prefix. When no suffix is found, the program goes to the logic rules, inserts the hyphen, and then returns to the suffix search to examine the syllable before the hyphen.
 4. Possible hyphens are inserted into the word provided that there
 - a) are at least five characters in the word
 - b) are at least two characters before the hyphen
 - c) are at least two characters after the hyphen
 - d) is at least one vowel in the syllable preceding the hyphen.
 5. The logic hyphenation consists of 8 rules. In the following
 - v indicates the presence of a vowel
 - c indicates the presence of a consonant.

- Rule 1: A syllable must contain at least one vowel. When a vowel is found the pointer is located at the vowel and the program steps on to Rule 2.
- Rule 2: Diphthongs may not be broken. Check for AA, EE, IE, OE, AI, UI, IJ, OO, AU, EU, OU. If the second vowel is a J go to Rule 2a. If vowel is a Y go to Rule 2b. If the first vowel is not part of a diphthong go to Rule 3.
- Rule 2a: Treat IJ as a single vowel. If IJ is found, handle this as a diphthong whose second letter is I and return to Rule 2. Else hyphenate the syllable as -J, and return to the suffix check. (IJ no longer handled in the 1986 algorithm).
- Rule 2b: If the vowel is Y, hyphenate as Y- and return to the suffix check.
- Rule 3: Hyphenate between vowels. If first character is a vowel, insert a hyphen as v-v, and return to the suffix check. If the first character is not a vowel go to Rule 4.
- Rule 4: Always break before Q or Z. If character before vowel is Q or Z, hyphenate as -Qv or -Zv, and return to the suffix check. If the consonant is not Q or Z go to Rule 5. (Taken out in the 1986 algorithm).
- Rule 5: Always break before a single consonant between vowels. If vcv the hyphenate as v-cv and return to the suffix check. If the character before the consonant is not a vowel go to Rule 6.
- Rule 6: Always break before consonant pairs, with exceptions. If ccv then hyphenate as -ccv unless it is found in the following list of exceptions.
 -SCHv else hyphenate as -CHv; -PHv; -THv (now excluded)
 T-Kv
 -BLv; -CLv; -FLv; -PLv; -SLv; -VLv
 -KNv; T-Nv
 vS-Pv; c-SPv; c-Pv
 -BRv; -CRv; -DRv; -FRv; -GRv
 -SKRv else hyphenate as K-Rv
 vS-PRv; c-SPRv else hyphenate as -PRv
 vS-TRv; c-STRv else hyphenate as -TRv
 -VRv
 vS-Tv; c-STv
 c-Tv
 T-Vv
- Rule 7: Always break between H and a consonant, with exceptions. Hyphenate as H-cv. No exceptions listed.
- Rule 8: Always break between two consonants, with exceptions. No exceptions listed.
- Rule 8a: Hyphenate as vc-cv
- Rule 8b: H-cv
- Rule 8c: vcc-cv
- Rule 8d: H-cccv
- Rule 8e: cc-ccv

Author Gee Quentin H

Name of thesis Automatic Hyphenation Of Afrikaans. 1987

PUBLISHER:

University of the Witwatersrand, Johannesburg

©2013

LEGAL NOTICES:

Copyright Notice: All materials on the University of the Witwatersrand, Johannesburg Library website are protected by South African copyright law and may not be distributed, transmitted, displayed, or otherwise published in any format, without the prior written permission of the copyright owner.

Disclaimer and Terms of Use: Provided that you maintain all copyright and other notices contained therein, you may download material (one machine readable copy and one print copy per page) for your personal and/or educational non-commercial use only.

The University of the Witwatersrand, Johannesburg, is not responsible for any errors or omissions and excludes any and all liability for any errors in or omissions from the information on the Library website.