3.3 Dictionary

The use of a hyphenation dictionary is common in most countries because it is the easiest to implement. The entire word list is stored on the computer disc with the correct hyphenation points marked. This also allows exceptions to be stored, such as words not to be hyphenated. Part of such an existing dictionary is illustrated below.

ge1bieds1hoof1kwar1tier ge1diend ge1digte ge1ding ge1dwee ge1dweep
gees1drif gees1drif1tig ge1kwas ge1kweel ge1kwes ge1kwets ge1kwyl
ge2nees1inrig1ting ge2weld1da1dige ge2weld1dadig1heid
ge1weld1da1dig1he1de ge1wiek ge1wig ge1wigte ge1wild ge1willde
ge1wisse giet1kuil glencairn gre2go1riaans me3de1plig1tige
me3de1plig1tiges me3de1plig1tig1heid me3diaan me3dies me3di1kus
me3dium mei1wyn mia2lie1wurm mie2li1tis mier1olie mo3der1nis1ties
mo3der1nis2tie2se

Here the numbers indicate preferred hyphenation points in descending priority order. For instance, if the hyphenation level is preset as 2 by the user in his document formatting, hyphenation will only take place at preference 2 and 1 positions; similarly, if he makes it 3, only preference 3 and below will be indicated.

This is very much a sledgehammer approach but is the one currently adopted in this country. Although easy to implement, it is probably the most expensive in the long run since it takes a large amount of storage, and can also be very slow due to the disc accessing required. We also note that the following words in the dictionary given to us contain spelling errors, and do not indicate all hyphenations that are acceptable to H.A.T. (Odendal, 1981).

ge2weld1da1di1ge, ge2weld1da1dig1heid, ge1weld1da1dig1he1de

The problem of speed can be overcome if the dictionary is used at the time a document is actually printed, when only those words needing the hyphenation need to be looked up by the program. Alternatively, if the philosophy is to apply it to every word, then it can be used at data capture stage when the speed of the operator is much slower than that of the computer. This would work as follows. While the user is entering the text, characters appear on the line in the order typed. At the end of the line word wrap usually occurs, which means that where the word is longer than the available *screen* space, the whole word is moved to the next line. The computer marks the space following the previous word with an invisible end-of-line character so that it can ignore irrelevant spaces during any subsequent changes to that text.

The advantage of the dictionary approach is that it is very easy to amend (assuming that access to the source is available) to enter or delete hyphenation points according to local rules. New words can be added such as proper names as mentioned earlier.

This approach has been successfully implemented in South Africa. IBM assembled a wordlist of the language and marked all the correct hyphenations in it together with endings and other features. This was then condensed into a binary tree/decision table system which is looked up when a line end decision is required while the user is entering or editing the text. In effect this is a complete dictionary.

The original text list was some 190 000 words, which would occupy some 96 floppy discs. The binary version fits nicely onto a 320 KB floppy disc and so is available for Displaywriter PCs (Malone, 1978).

3.4 Semi-manual

This method is frequently implemented in the personal computer software, which has been outlined in section 2.5, whereby every time the computer nears the end of a line, it displays a message asking the user to provide an acceptable hyphenation point. This, of course, exactly mirrors the old mechanical approach of the Linotype and Monotype machinery. Better software implementations are available which suggest several possible break points to the user, but these approaches appear to be based on English syllable rules and so are not very accurate for Afrikaans.

The semi-manual method is only realistic in document processing where the characters all have the same width, and therefore the line length can be preset. Where variable width characters are used, as in typesetting, it stops the process every time a decision is required. This is slow and reduces productivity.

3.5 Expert Systems

A new discipline within computing science is that of Expert Systems. The objective is to explore the way in which the human mind tackles problems or situations, mirror that thinking process where possible, and provide a reasoning system capable of accumulating data and drawing conclusions from it.

Typically a framework (shell) is used to store and manipulate data about the problem, and to store inferences drawn from the data. In this it is initially helped by an informed user (expert), but subsequent information has the effect of making the system decide whether it can fit this into its knowledge of the situation, or modify its current set of inferences, or whether it must ask for guidance on how to interpret this new information. Some systems allow the user to request the working methodology, which is the set of inferences it has used, why some of these have been rejected or accepted, and what the next goals and sub-goals are.

This approach to hyphenation has not yet been tried but could be very effective for the future where the system can modify its own set of rules as it 'learns' i.e. is taught what is acceptable and what is unacceptable. Hyphenation would be a very good test for an Expert System since the problem is well defined, and limited in scope, and the result can be easily tested.

3.6 Decision Tables

Decision Table systems are powerful tools for condensing decisions that are based on logic. The methodology requires a set of structured matches to be performed, and for each match found, an action is provided. This action may be a further matching process within the current table, or it may route the query to another table, or it may deliver a result.

We believe that such an approach is likely to produce the best solution to the hyphenation problem, although not an elegant one.

A valid implementation would start with the dictionary outlined in 3.3 above. The dictionary must be complete, the words correctly spelt, and all the valid hyphenation points must be indicated. Using this as input data, a set of decision tables can be produced automatically by many easily available software packages.

The resulting set of tables is very fast in operation, uses a small amount of main memory, but is difficult to amend and can only be altered by re-generating the whole set of tables from the original dictionary.

To show how this works, in the following example we take a short list of words and assume that they are correctly spelt and have all their syllables correctly hyphenated.

ge-bieds-hoof-kwar-tier
ge-diend
ge-digte
ge-ding
ge-dwee
ge-dweep
gees-drif
gees-drif-tig
ge-kwas
ge-kweel
ge-kwes
ge-kwets
ge-kwyl
ge-nees-inrig-ting
ge-weld-da-dige
ge-weld-da-dig-heid
ge-weld-da-dig-he-de
ge-wiek
ge-wig
ge-wigte
ge-wild
ge-wil-de
ge-wisse
giet-kuil
glen-cairn
gre-go-riaans
me-dium
mei-wyn
mie-lie-wurm
mie-li-tis
mier-olie
mo-der-nis-ties
mo-der-nis-tie-se
me-de-plig-tige
me-de-plig-tiges
me-de-plig-tig-heid
me-diaan
me-dies
me-die-se
me-di-kus

From this information decision tables are constructed using the following conventions.

N	do not break after the first letter designated
Y	insert a hyphen after the first letter designated
digit	search the table so numbered for a further test
.	does not occur; take no action
$	indicates an end of word symbol e.g. space, comma, quad

TABLE 1

First letter	Second Letter A	B	C	D	E	F	G	H	I	K	L	M	N	O	P	R	S	T	U	W	Y	$
A	N	.	.	Y	.	.	.	.	N	.	.	.	N	.	.	N	N	.	.	.	.	.
B	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	.
C	N	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D	N	.	.	Y	N	.	.	.	N	.	.	.	.	.	.	N	N	.	.	N	.	N
E	.	N	.	3	N	.	Y	.	N	3	3	.	3	.	3	N	N	N	.	Y	.	N
F	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	Y	.	.	.	N
G	.	.	.	.	.	.	.	Y	N	.	.	.	.	N	.	N	.	2	.	.	.	N
H	.	.	.	.	N	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	.
I	N	.	.	N	N	N	N	.	.	Y	N	.	N	.	N	.	N	Y	N	Y	.	.
K	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	N	N	.	N
L	.	.	.	4	N	.	.	.	N	.	.	.	.	.	.	.	.	.	.	.	.	N
M	.	.	.	.	N	.	.	.	N	.	.	.	.	N	.	.	.	.	.	.	.	N
N	.	.	Y	N	N	.	N	.	N	.	.	.	.	.	.	N	N	.	.	.	.	N
O	.	.	.	Y	.	N	.	.	.	.	N	.	.	N	.	Y	.	.	.	.	.	.
P	.	.	.	.	.	.	.	.	.	.	N	.	.	.	.	.	.	.	.	.	.	N
R	.	.	.	.	N	.	.	.	N	.	.	N	5	Y	.	.	.	Y	.	.	.	N
S	.	.	.	Y	.	.	.	Y	Y	.	.	.	.	.	.	.	N	Y	.	.	.	N
T	.	.	.	.	N	.	.	.	N	Y	.	.	.	.	.	.	N	.	.	.	.	.
U	.	.	.	.	.	.	.	.	N	.	.	N	.	.	.	N	N	.	.	.	.	.
W	N	.	.	.	N	.	.	.	N	.	.	.	.	.	.	.	.	.	N	.	N	.
Y	.	.	.	.	.	.	.	.	.	.	N	.	N	.	.	.	.	.	.	.	.	.

TABLE 2

First group	Followed by E	I	Action
GT	N	Y	break after G

TABLE 3

First group	Followed by C	D	E	I	L	S	W	$	Action
ED	.	.	Y	Y	.	N	Y	.	break after E
EK	.	.	.	.	.	.	Y	N	break after E
EL	.	N	.	Y	.	.	.	N	break after E
EN	N	N	Y	.	.	.	.	.	break after E
EP	.	.	.	.	Y	.	.	N	break after E

TABLE 4

First group	Followed by D	E	$	Action
LD	N	Y	N	break after L

TABLE 5

First group	Followed by I	$	Action
RN	Y	N	break after R

Suppose we now decide that an extra hyphenation point is acceptable in the word ME-DIAAN, and the entry in the list is changed to ME-DI-AAN. The 'N' entry in Table 1 would now be incorrect, because we have both ME-DI-AAN and GRE-GO-RIAANS which have different IA breaks. Therefore a re-generation of the decision table would produce an additional Table which would differentiate between these two words. In this simple example it has only been necessary to provide additional decision tables with 'followed by' entries. With a more extensive list, there will be tables with 'preceded by' entries. In order to optimise the entries some intricate work would have to be performed by hand because it may be easier for the program to check EN-SASIE rather than DISPEN-S. This optimisation can obviously better be performed visually, but such a manual operation would be tedious and errorprone.

Chapter 4 APPROACH ADOPTED

In order to produce an acceptable system which will work on a PC, an approach must be adopted which uses little storage space, has a fast accessing time, and is easy to amend. It is clear that the use of an algorithm will satisfy the first two conditions, while the addition of an exception dictionary will cater for both the exceptions to the algorithm and idiosyncratic exceptions, and can be updated with local changes easily. The rules can be deduced by inspection of suitable word lists and reference to authorities. It is also easy to check the results against such sources.

4.1 Morphotactics

Since a set of rules will have to be set up, the rules should mirror real phonetic features and not, as in the decision table approach, some apparently arbitrary concatenation of symbols. The rules should be easy to modify as further research is performed on the outstanding problem areas.

Thus research has to be undertaken to specify all syllables in Afrikaans, and the points at which compound words are glued together. This is a branch of morphology called **morphotactics**.

The term **syllable** is used by the linguistic scientists in a number of ways. The definition which concerns us is the phonological one.

> "Phonetically the term syllable has frequently been used to refer to a sequence of speech sounds having a maximum or peak of inherent sonority (that is apart from factors such as stress and voice pitch) between two minima of sonority. ... as part of the phonological level of analysis, ... the syllable refers to a number of different sequences of consonants and vowels, together with other features such as length and stress, or to single consonants or vowels, which ... are suitably considered as a unitary group for further analysis.
> Phonetically it is often hard to decide to which syllable an intervocalic consonant belongs, ... and such consonants may be ... ambisyllabic. ... in general single intervocalic consonants within a word are assigned to the following vowel." (Robins, 1964, pp137-138).

We therefore decided to investigate this approach in detail.

4.2 Algorithm + Exception Dictionary

Starting with an algorithm, we list in the dictionary all those words which do not satisfy the rules as known, whether because they are real exceptions, foreign words, or idiosyncratic occurrences. This list is then searched first and words not caught during that process are then submitted to the algorithm. Thus our objective is to make the dictionary as small as possible.

4.3 Affixes vs Syllables

The first question that arises when attacking the problem of analysing syllables for computerisation is: Should affixes be treated differently or just in the course of normal syllable identification? The latter is most convenient and elegant and needs the same set of rules throughout. In other words, where the affix would be broken correctly using the syllabification rules, do we need a special affix search? Most working algorithms in European languages use the fact that Prefixes and Suffixes exist and can be used to shorten the searching methodology. This can be naïve and simple (break after AAN-) or it can be complex and detailed (break after AG- if not followed by A, E, I, O, R, or U). Exceptions can always be loaded into the dictionary if not specifically coded in the algorithm.

We note that while English manages with most word structures in the form

Prefix Root Suffix

where two prefixes and two suffixes (RE-IN-CAR-NA-TIONS) may be affixed, German and Afrikaans are agglutinative languages and so many compound words exist which provide word structures in the form

Prefix Root Suffix Prefix Root Suffix

with multiple prefixes and suffixes in all positions, for example

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

We accept the proposition that affixes should be treated separately since they provide a quick initial analysis, may provide adequate break points, and help medial problems in compound words.

Our aim is to provide a computer procedure which returns to the calling program the position of the character closest to, and to the left of, the virtual line end. That is, the line end less the width of a possible hyphen. It may return NO BREAK also.

Chapter 5 PREFIX RESEARCH

It is clear that in Afrikaans the elements to be tackled are prefixes, suffixes and vowel/consonant combinations. In addition, we must ensure that the boundaries of compound words are addressed, if not already covered by the vowel/consonant research.

> "Affix is a useful general term [illegible] the recurrent formative morphemes of words other than roots, but affixes n[illegible] divided ... according to the position they occupy in relation to the root morpheme: *prefix*, *infix* and *suffix*." (Robins, 1964, p 210).

However, there are no infixes in Afrikaans.

5.1 Prefixes

What do we mean by a 'Prefix' ? A **prefix** consists of those letters that precede the root word, and cannot be used on their own.

One approach is to use common *beginnings* of words, and consider these as prefixes also (Forsyth, 1986), but this brings us back to the Affix vs Syllable approach which we have rejected for the purpose of this research. Most of these *beginning* features comprise meaningful words on their own. A sample list of those suggested by Forsyth are given below, but could comprise all roots in the language.

Begin cluster	Exceptions
HEER-	
MELD-	
MYN-	MYNENT, MYNER, MYNERSYDS
PAD-	PADIE
SKOON-	
SPOOR-	
TWEE-	TWEË
TYD-	TYDELIK, TYDELOOS, TYDENS, TYDIG, TYDIGHEID, TYDING

VUUR-

WATER- WATERSNOOD

WERK- WERKER, WERKERSBOND, WERKERIG, WERKIE, WERKING, WERKSDAG, WERKSMAN, WERKSWINKEL

WOORD-

WOOR-DE-

However, if we eliminate all but popular prefixes as defined in AW&S (1964), then we are only left with the following, which is too small a set to be practically useful in identifying and detaching prefixes effectively.

BE-, DE-, GE-, HER-, ON-, ONT-, TE-, VER-

Taking the major list in Kempen (1977), together with selected items from Pretorius (1972) and Marais (1970) and amalgamating these, we arrive at a workable set of prefixes. We then choose a middle course of action by defining a clear set of classes and use reasonable discretion on the inclusion of elements. Some of those listed below could occur in more than one category because they can have more than one meaning, for example. We only mention the element once; having included it, the reason why we did so is irrelevant as we are thereafter only concerned with orthography.

1. Real prefixes

AARTS-, BE-, DE-, DER-, ER-, GE-, HER-, MIS-, OER-, ON-, ONT-, TE-, VER-, WAN-

2. Parts of compound verbs

AF-, IN-, OM-, OOP-, OP-, SAAM-, SAME-, TOE-, UIT-, VAS-, WEG-

3. Adverbs of Place/Size/Time

AAN-, AGTER-, BY-, DAAR-, DEUR-, HIER-, NEER-, NOORD-, ONDER-, OOR-, OOS-, OOSTE-, OPPER-, OWER-, SUID-, SUIDE-, TEEN-, TERUG-, TUSSEN-, UITER-, VAN-, VOL-, VOOR-, VOORT-, VRY-, WAAR-, WEDER-, WEER-, WES-, WESTE-

4. Greek/Latin prefixes (after Marais, 1970)

A-, AB-, AD-, ALLO-, AMBI-, AMFI-, ANA-, ANTE-, ANTI-, APO-, ARGE-, ARGI-, BI-, BIO-, BRAGI-, DES-, DIA-, DIS-, EKS-, EKWI-, ENDO-, ETNO-, GEO-, HEKSA-, HEMI-, HETERO-, HIPER-, HIPO-, HOMO-, IDIO-, INFRA-, INTER-, INTRA-, ISO-, KATA-, MAKRO-, META-, MIKRO-, MONO-, MULTI-, ORTO-, OUTO-, PARA-, PERI-, POLI-, PRE-, PRO-, RE-, SIM-, SIN-, SUB-, SUPER-, TELE-, TRANS-, ULTRA-

5. Other obvious elements

EEN-, FA-*, KAR-*, KE-*, KER-*, MEDE-, MEER-, SELF-, WEL-

Although A- is included as a prefix in group 4, according to typographic rules it may not be left on a line on its own and is therefore ignored at the beginning of words. Should it occur as a prefix within a compound word, practice also dictates that it cannot end a line.

Using this scheme, the prefix elements are listed below in alphabetical order together with all the words that do not conform to the appropriate rule for breaking each prefix. These exceptions were determined by inspection of dictionary authorities including Odendal (1981), Bresler (1978), Kritzinger (1968), Bosman (1984).

The conventions used are as follows. The symbol '-' indicates an acceptable hyphenation point at that position in the word. For example, BE- means that a hyphen is always acceptable after the BE prefix. Words are listed in capitals and mainly comprise the *minimum* number of letters required to effectively discriminate between alternatives. Thus exceptions listed under BE- for example, include BEU, which means that in all words starting with BEU, the BE may not be hyphenated.

The prefixes cannot necessarily be used in this order because, for example, a test for DE will produce DE-, whereas the word may be DERDE which breaks DER-DE. Thus DER must be tested before DE.

---

* 'n Vreemde Prefiks - Anna E Coetzee in Sinclair A J L (red.) : *G S Nienaber - 'n Huldeblyk,* UWK Drukkery, Kaapstad, 1983.

Prefix	Non-conforming words
AAN-	AAND (except for AAN-DAG, AAN-DEEL, AAN-DENKING, AAN-DIEN, AAN-DIK, AAN-DOEN, AAN-DRA, AAN-DRING, AAN-DRUK, AAN-DRYF, AAN-DUI, AAN-DURF)
AARTS-	none
AB-	ABA (except for AB-AKS), ABE, ABI, ABO, ABRA, ABU
AD-	ADA, ADE, ADI, ADO, ADR, ADU
AF-	AFRIC, AFRIK, AFERESE, AFONIE, AFASIE
AG-TER-	none
AL-LO-	none
AM-BI-	none
AM-FI-	none
ANA-	ANALF, ANAM, ANAR, ANAS, ANATT
ANTE-	ANTENNE
ANTI-	ANTIEK
APO-	APOS
AR-GE-	ARGEN
AR-GI-	ARGIE
BE-	BEATNIK, BEDB, BEDD, BEDJ, BEDK, BEDL, BEDS, BEDT, BEDV, BEDW (except for BE-DWANG, BE DWELM, BE-DWING, BE-DWONG), BEE (except for BE-ËI, BE-ËL), BEF, BEIAARD, BEIDE, BEIER, BEIGE, BEITS, BEKAF, BEKFL, BEKK, BEKPR, BEKRIEM, BEKS, BEKV, BELBREI, BELG, BELHAM, BELK, BELL, BELR, BELS, BELT, BELV, BENDE, BENG, BENJ, BENN, BENS, BENT, BERB, BERD, BERG, BERK, BERL, BERR, BERS, BESKIN, BESS, BESTIAL, BESWIL, BETSJ, BETWE, BEU
BIO-	BIOPSIE
BI-	BIBB, BIDD, BIDS, BIE, BIGG, BIL (except for BI-LA), BIN (except for BI-NÊ, BI-NO), BIS (except for BI-SA, BI-SE, BI-SO), BIT (except for BI-TU)
BRA-GI-	none
BY-	BYL (except for BY-LAE, BY-LANGE, BY-LÊ, BY-LEER, BY-LEGGING, BY-LOOP, BY-LYN), BYSTER, BYT, (except for BY-TEL, BY-TEND, BY-TER, BY-TERIG, BY-TOON, BY-TREK)
DAAR-	none
DER-	DERI, DERMAF, DERMB, DERMK, DERMO, DERMS, DERO
DES-	DESEM, DESEN, DESER, DESI (except for DES-ILL, DES-INF, DES-INT), DESKR, DESOL, DESTR
DEUR-	DEURE