AUTOMATIC HYPHENATION OF AFRIKAANS

Quintin H Gee
November 1987

A Dissertation Submitted to the Faculty of Science
University of the Witwatersrand, Johannesburg
for the Degree of Master of Science

AUTOMATIC HYPHENATION OF AFRIKAANS

Quintin Gee, University of the Witwatersrand

When typed text will not exactly fill a line, it is necessary to hyphenate the last word correctly. When this function is controlled by computer program, we must supply that program with a set of rules pertaining to the language being used.

For Afrikaans, linguistic hyphenation rules such as 'Break between meaningful components' cannot be used by computers without encoding word pronounciation and semantics. Where continuous text is being processed, for example in word processing or typesetting, such an approach is both complex and unnecessary. We need rules that relate to the orthography. For example 'Always split doubled consonants', but inevitably there are exceptions to this.

One approach to supplying correct hyphenation points to computer programs is to store a whole Afrikaans dictionary containing the hyphens already positioned manually. This does not accommodate the creation of new compound words and thus cannot be said to be a general solution, although it does satisfy the requirements of all common words. For personal computers, such a dictionary is also too expensive an approach to undertake currently, and an algorithmic solution is preferable, which uses a dictionary only to contain idiosyncrasies such as imported words and proper names.

This research describes the orthographic formation of syllables in Afrikaans and analyses them to determine how to recognise syllable boundaries, multiple prefixes and multiple suffixes. These three elements are then combined in a computer procedure to locate hyphenable positions in a word. Words not following these rules are coded in internal tables. Examples are given of applying this to a word list and to continuous text.

AFRIKAANSE LETTERGREEPVERDELING OP REKENAAR

Quintin Gee, Universiteit van die Witwatersrand

Wanneer getikte teks nie in 'n reël pas nie, is dit nodig om die laaste woord in die reël korrek te verdeel. Waar hierdie funksie deur 'n rekenaarprogram beheer word, moet 'n program van 'n stel reëls in die besondere taal voorsien word.

Afrikaanse taalkundige reëls vir lettergreepverdeling soos 'Verdeel tussen twee betekenisvolle komponente' kan nie deur rekenaars gebruik word sonder om die uitspraak en betekenis van woorde te enkodeer nie. Waar deurlopende teks geprosesseer word, byvoorbeeld in woordverwerking of setwerk, is so 'n benadering ingewikkeld en onnodig. Ons het reëls nodig wat betrekking het op die ortografie, byvoorbeeld: 'Verdeel altyd tussen dubbele konsonante', maar ook hier is onvermydelike uitsonderinge.

Een benadering tot die verskaffing van lettergreepfeite aan rekenaarprogramme is om 'n Afrikaanse woordeboek, met die lettergreepverdelings daarin aangebring, in sy geheel te stoor. Alhoewel hierdie benadering sorg dat alle gewone woorde sc geakkomodeer kan word, kan dit nie dien as 'n algemene oplossing nie, want nuwe samestellings wat geskep word, sal probleme veroorsaak. So 'n woordeboek is 'n te duur benadering om huidiglik vir persoonlike rekenaars te onderneem. 'n Algoritmiese oplossing, met 'n woordeboek wat uitsonderings en spesiale gevalle, soos leenwoorde en eiename bevat, is verkieslik.

Hierdie navorsing omskrywe die ortografiese vorme van lettergrepe in Afrikaans en analiseer lettergrepe ten einde lettergreepgrense en veelvoudige voor- en agtervoegsels te herken. Laasgenoemde drie elemente word gekombineer in 'n rekenaarprosedure om posisies van lettergreepverdelings in 'n woord te bepaal. Woorde wat nie hierdie reëls volg nie, word in interne lyste gekodeer. Voorbeelde van toepassings op sowel 'n woordelys as aaneenlopende teks word gegee.

DECLARATION

I declare that this dissertation is my own, unaided work. It is being submitted for the degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

21st day of November, 1987

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to my supervisors, Professor J M Bishop of the Computer Science Department and Mrs A E Coetzee of the Afrikaans en Nederlands Departement, for their encouragement and assistance throughout and for their patience in dealing with a multi-disciplinary approach.

My sincere gratitude is also due to the Managing Director and staff of Atex Systems (Pty) Ltd, part of the Kodak group, for making their computing facilities available to me. Without use of their high-speed text searching equipment, and access to a source list of 100 000 Afrikaans words, this work would not have been completed.

Finally, I would like to dedicate this work to my wife Sonÿa, who supported me through a difficult journey.

Aan Sonÿa, wat die moeilike reis met my meegemaak het.

Chapter 1 INTRODUCTION

For the past 400 years, printers using movable type have had two problems to overcome when fitting text into their rectangular formes: **justification** and **hyphenation**. The first was solved by trying to make each interword space the same width. To do this printers either inserted identical spaces between words throughout the manuscript and then had to fill up the ends of lines with a variable amount of space, or they put a variable amount of space between each word, so that there was no slack at the end of a line. Until the advent of mechanical typesetting, justification was performed manually using a standard set of 5 space widths; since then, variable spacing has been used.

The hyphenation of words in a given language is a relatively easy process for the mother-tongue typist since a word can be enunciated syllable by syllable, and these inter-syllable boundaries provide natural break points. The same syllabic technique applies to all European languages and can also be extended to the hyphenation of other languages whose writing system is alphabetic. Before computers were used in the printing industry, typesetting was a manual or semi-manual operation. The keyboarding device would warn the operator when the end of the line was approaching, much as a manual typewriter rings a bell. At that juncture the operator had to make a decision on whether to insert a hyphen in the last word, and where one would be appropriate.

The introduction of computers has revolutionised the printing industry in many areas. Computerised typesetting operates in the following manner. Manuscript text may be keyboarded directly into the typesetter by an operator, or be received by the typesetting computer from another computer by means of data transmission along a telephone line. It may also be brought physically on some form of computer-readable magnetic medium such as magnetic tape, floppy disc, or cassette. A computer program, commonly called a **composition program**, then formats the text according to **justification rules**, and translates from the computer's data representation into the typesetter's data representation. This machine then produces photographic film, a positive paper image, or even a printing plate directly.

One of the rules within a justification procedure concerns the variable spacing mentioned above, and another the automatic hyphenation of words. However, the computer can only carry out a set of instructions and has no understanding of the meaning of the words or their pronunciation; all it can operate on is a string of symbols. This means that rules such as 'Break between meaningful components' (AW&S, 1964) are not very useful, although linguistically accurate. An example of a rule more acceptable for the computer would be 'Always split doubled consonants', but inevitably there are exceptions to this.

Publishers and printers in South Africa need to correctly hyphenate many languages. Hyphenation rules for many European languages are available, notably English (Knuth, 1979; Ocker, 1975), French (Hart's Rules for Compositors and Readers at the University Press, Oxford, 1967), German (Martin, 1964 & 1965; Bithell, 1952), Spanish (Mañas, 1987), and Dutch (Forsyth, 1986). While the rules for hyphenating English are well understood now, they are not widely implemented outside the printing industry. Little formal work has been undertaken in South Africa on the hyphenation of other languages except for North Sotho (Gee and Price, 1988), although the handling of African languages is widely implemented in the printing industry and has been moderately successful using rules-of-thumb. In particular, Afrikaans has not been properly addressed to achieve its correct and efficient hyphenation, nor do most computers handle its accents properly.

As the use of personal computers becomes more widespread, the problem of handling Afrikaans text correctly intensifies. Indeed, most word processing software and hardware for sale in this country ignore both accents and hyphenation.

Since the cost of computer disc storage has dropped dramatically in the last 10 years, *sledgehammer* tactics to hyphenate Afrikaans have most frequently been adopted. Thus, "a whole Afrikaans dictionary containing 190 000 words has been stored which contains all the hyphens already positioned manually", (Malone,1987). This approach inevitably ignores the problem of creating new compound words and thus cannot be said to be a general solution, although it does work in practice for all common words. While the speed of accessing information stored on the computer's discs has not increased radically during that time, this method is nevertheless a viable one for large computers which can continue with other tasks while searching takes place.

However, for small computers this is certainly too expensive an approach to undertake currently, and an algorithmic solution would be preferable, relying only on a dictionary to contain idiosyncrasies such as imported words and proper names.

Source material was a bit sparse. A few theses exist which take as their subject the prefixes and suffixes which occur in Afrikaans. Some computer literature on the subject of hyphenation is available e.g. one paper in Engineering (Ocker, 1975), one about Spanish (Mañas, 1987), some rules about English (Knuth, 1979).

The work detailed in this thesis therefore set out to

- Undertake research into the orthographic formation of syllables in Afrikaans
- Analyse the syllables to determine how to recognise syllable breaks and breaks between compound words
- Condense this analysis into a set of rules for practical application to personal and other computers
- Test the set of rules in a practical environment
- Publish the results for the benefit of South African printers, publishers, and computer suppliers.

Chapter 2 PURPOSES

2.1 Orthographic research

Our investigation has revealed that the rules for pronunciation of Afrikaans are well-known. The mapping of the way words are written, the **orthographic** system, to the pronunciation rules, is consistent (AW&S, 1964). For example, the letters written EI or Y are always pronounced [əi], and the characters F and V are pronounced [f] when occurring at the beginning of a word. This mapping may not always be one-to-one. This arises historically from the different origins of the words, for example we have:

FEIL	[fəil]
VEIL	[fəil]
VYL	[fəil]

In the early discussions on the language, rules were formulated that ensured consistency of pronunciation while maintaining the historical ties with earlier orthographies (AW&S, 1964). The rules governing the reverse mapping, that of the pronunciation to its orthography, are also consistent, the examples given above being few in the language and generally known by a mother tongue speaker.

The objective of our new work has been to codify the mapping of morphemes to spelling for computer purposes, and to identify any exceptions in the procedure, so that, if required a *complete* set of rules can be tabulated.

One of the difficulties in setting a realistic objective is 'What constitutes correct hyphenation?' There are words that are **homographs** and that break differently in different contexts (GEK-LIK, GE-KLIK; HOU-TJIE, HOUT-JIE). There are proper names which must never be broken or which must be broken in non-standard ways (VILJOEN). Then there are words that include punctuation marks and normal hyphens (VOLKS- EN GEMEENSKAPSGESONDHEID). There are imported words that contain 'foreign' spellings (CAMOUFLAGE). This research tackles these problems and suggests practical solutions for them that avoids extensive literary analysis.