



**MORPHOLOGY AND STRESS
IN A RULE-BASED GRAPHEME-TO-PHONEME CONVERSION SYSTEM FOR DUTCH**

Egon Berendsen* and Jan Don^

SUMMARY

This paper deals with the interaction between morphology and stress assignment in a rule-based grapheme-to-phoneme conversion system for Dutch which is being developed at the Universities of Utrecht and Leyden and at the Institute of Perception Research at Eindhoven (see also ref 1 and ref 5). In order to establish the interaction between morphology and stress assignment we will deal with some aspects of our rule system, present a survey of the relevant linguistic data in Dutch and show how the former accounts for the latter.

1. THE SYSTEM

Our grapheme-to-phoneme conversion system is rule-based. The leading idea has been that linguistic knowledge has to be separated from computational knowledge. This implies that an independent linguistic grammar has to be implemented in a computer system. As a result the system is user-friendly for a linguist, because it is not necessary for him to have considerable computer knowledge. Furthermore, additions and changes in the linguistic grammar can be made interactively, because the system does not have long compilation times. Finally, its user-friendliness is increased by the rule-format chosen. The linguistic rules used are of the well-known Sound Pattern of English type (ref 2), as is exemplified in (1).

Focus -> Change / Left Context _ Right Context (1)
where focus, change and context may be: grapheme(s),
phoneme(s), grapheme feature(s), phoneme feature(s), empty

It turned out to be necessary to extend this rule format in the following way: a) it should be capable of negating elements or groups of elements which is indicated by a quote; b) it should allow for the use of so-called global rules, in such a way that one can refer back to graphemic information on the phonemic level; c) it should be possible to use definitions rather than frequently used sequences; d) it should be possible to require that two or more conditions are met at the same time in one position, so-called coordinations, indicated by the symbol +.

To obtain information about the relation between graphemes and phonemes our system recognizes both a grapheme level and a phoneme level. Roughly speaking, a phoneme is assigned to each grapheme or group of graphemes, as in (2). This two level approach offers the possibility to use the global rules referred to above.

grapheme level	a	b	c	d	e	i	(2)
						/	
phoneme level	A	B	K	D	E	I	

The graphemic sequence to be converted is scanned sequentially. This can be done from left to right or from right to left. Furthermore, if the

Departments of *^Linguistics and *Computers & Humanities,
University of Utrecht, Trans 14, 3512 JK Utrecht, The Netherlands

grapheme under consideration is for example an a, only the rules which take an a as their input will be considered, with the additional limitation that if a rule matches, the process will be executed and the next grapheme in the sequence will be considered. As a consequence, the output of one rule cannot immediately be the input of another rule. However, the output of one scan through the string can be the input for another, due to the modular nature of our system. Three kinds of modules are distinguished: a) input graphemes/output graphemes (gragra); b) input graphemes/output phonemes (grafon); c) input phonemes/output phonemes (fonfon). There may be as many gragra and fonfon modules as one needs. However, there may only be one grafon module because in that module every grapheme has to be linked to the phoneme level, whereas in the other modules a grapheme or phoneme to which no rule has applied, is simply copied to the output.

2. MAIN STRESS IN DUTCH WORDS: SOME OBSERVATIONS

In Dutch monomorphemic words, stress is always located on one of the final three syllables. As theoretical research has shown (cf. for example ref 3), the position of stress in these three syllables depends on the weight of their syllable-rimes. Thus, we arrive at the survey of the unmarked stress-positions in (3). There VV indicates a long vowel, V indicates a short vowel, and C a consonant. The same holds for specific vowels and consonants. X is a variable ranging over consonants. A number after a C means: at least that number of consonants. Finally, \$ represents a word-boundary, i.e. a blank.

- I Ultimate stress in words ending in rimes of the form: (3)
- (i) VVC1 or VC2: kanáal (canal), stochást (stochast)
 - (ii) UU, EE, or diphtongue: paraplú (umbrella), karwéi (job)
 - (iii) ET, ES, EL, IN, IL: adrés (address), kapél (chapel)
 - (iv) \$-VX: hónd (dog), vlá (custard)
- II Penultimate stress in words ending in rimes of the form:
- (i) VC-VV: agénda (agenda), embárgo (embargo)
 - (ii) VC-VC: eléctron (electron), abórtus (abortion)
 - (iii) VV-VV: aréna (arena), albíno (albino)
 - (iv) VX-∂X: parádē (parade), komkómmēr (cucumber)
 - (v) \$-VV-VC: ádam (Adam), fákir (fakir)
- III Antepenultimate stress in words ending in rimes of the form:
- (i) VX-VV-VC: álmanak (almanac), ánanas (pine-apple)

Polymorphemic words in Dutch can be derived by prefixation, suffixation or compounding. We will not go into matters of prefixation here. However, we will be dealing with suffixation and compounding.

With respect to stress assignment, four types of suffixation can be discerned, examples of which are presented in (4). Words derived by the first two types of suffixation behave with respect to stress assignment as if they were monomorphemic (see ref 4). Romein, for example, is of the monomorphemic type I(i). Words ending in a stress-bearing suffix in fact behave as if they belong to the monomorphemic types I(ii) or I(iii). However, in words derived by the last two types of suffixation the suffixes do count as separate entities with respect to stress assignment.

- | | | |
|------------------------------------|--------------------------------------|-----|
| a Roman suffixation | b Stress-bearing suffixation | (4) |
| Róme (Rome) - Rom-éin (Roman) | kóning (king) - koning-ín (queen) | |
| c Stress-neutral suffixation | d Stress-attracting suffixation | |
| éng (tight) - éng-heid (tightness) | próza (prose) - prozá-isch (prosaic) | |

The stress-neutral suffixes do not influence the main stress position of the stem. Thus, these suffixes have to be ignored so that the stress assignment rules can be applied only to the stem. The stress-attracting suffixes, on the other hand, trigger a shift of the main stress position in the stem to the first stressable (non-schwa) vowel to the left of the suffix.

With respect to compounds, the following observation can be made. Compound stress is usually located in the left-hand part of the compound on the same vowel that bears main stress if it is used in isolation. This is illustrated in (5a). However, if a stress-attracting suffix is a morpheme of the right-hand part of a compound, main stress is located on the first stressable vowel to the left of this suffix, as is illustrated in (5b).

a vóet - bál - vóetbal b díenst - plícht - dienstplichtig (5)
 (foot) (ball) (football) (service) (duty) (serviceable)

3. MORPHOLOGY AND STRESS IN OUR GRAPHEME-TO-PHONEME CONVERSION SYSTEM

To arrive at stress assignment in Dutch words, we showed in section 2 that it is necessary to split up words in some of its constituting morphemes. After that has happened our system will convert the split-up graphemic representation into a phonemic representation to which the stress assignment rules will be applied.

Since words derived with Roman suffixes or stress-bearing suffixes behave like monomorphemic words with respect to stress assignment, it is not necessary to recognize these suffixes. However, it is necessary to recognize stress neutral and stress attracting suffixes and parts of compounds and we insert boundary-markers (% , & and # respectively) to indicate them.

Generally speaking, Dutch spelling does not have any indication as to the word's morphological structure. Thus, to recognize suffixes and parts of compounds we can only rely on the word's graphemic information. For suffixes this seems rather easy, since the number of suffixes is limited. One simply strips off those sequences of graphemes which represent suffixes. However, there are cases in which it is not so simple. Consider for example the suffix -en (plural present tense, infinitive or plural noun), which is stress neutral. If we state our rule as in

$$e,n \rightarrow \% , e,n / \text{voc,cons0} _ \langle -\text{segm} \rangle \quad (6)$$

we correctly strip the suffix in examples as **benen** (legs) and **kopen** (to buy), but incorrectly in **grondpen** (ground pin). The graphemic sequence **dp** does not occur at word-edge and because of that the sequence **en** cannot be a suffix here. Thus we have to include this information in our rule as a negative condition in the left-hand environment.

Of course, compounds cannot be treated in the same way since the number of words is not limited, as is the number of suffixes. For these cases, we look for graphemic indications that a compound-boundary must be inserted. For example, sequences of a vowel, a consonant, an e, a consonant and a vowel normally represent a compound, as in **eikeboom** (oak-tree) and **keuzecursus** (optional course). This is expressed in rule (7).

$$e \rightarrow e, \# / \text{voc,cons} _ \text{cons,voc} \quad (7)$$

However, this rule overgeneralizes, in particular in cases with Roman suffixes. Rule (7) will incorrectly insert a boundary in for example **literatuur** (literature) and **fonetiek** (phonetics). These boundaries will

be deleted again in the last morphological module in our system. At present our system has three morphological modules, all of the gragra-type. The first primarily deals with suffixation and compounding and goes from right to left through the string. The second primarily deals with prefixation and the third is mainly an escape module in which overgeneralized boundaries are deleted and exceptional boundaries are inserted. In the latter two modules the graphemic string is scanned from left to right.

After the graphemic string has been converted into a phonemic string, it enters the stress assignment modules. Stress assignment is performed in a two-step procedure. In the first stress-module, the phonemic input string is being scanned from right to left and the rules for monomorphemic stress assign secondary stress. These rules express the observations made in (3). In this module, the morphological information is available in the form of boundaries. Compound-boundaries (#) are being handled as word-boundaries, which is to say that the rules for monomorphemic stress assignment apply to all parts of a compound. Stress neutral boundaries (%) count as a word-boundary in the right context of the rule, but not in the left context. In this way stress assignment on the suffix itself is blocked and will only be applied to the stem. The second step of the stress assigning procedure amounts to the heightening of the secondary stresses in two cases: in the left-hand part of a compound and in non-compounds. Both these cases can be captured by way of a rule that recognizes 'hard' word-boundaries, i.e. blanks. This rule is represented in (8), where \$ represents a blank. The right-hand context of this rule is needed to account for stress assignment in words with stress-attracting suffixes. In these words, primary stress has already been assigned in the first stress module to the first stressable vowel before the stress attracting boundary &. Rule (8) is not applicable here because it excludes a primary stress in its right context. Derivations now run as in (9), where Q is a velar nasal and C is a schwa.

$\langle 2\text{stress} \rangle \rightarrow \langle 1\text{stress} \rangle / \$, (\langle +\text{segm} \rangle) 0 _ '[Y, \langle 1\text{stress} \rangle]$ (8)
 where Y is a variable ranging over segments and
 word-internal boundaries

input:	romein	rilling	voetbal	fonetiek	dienstplichtig	(9)
morph 1/2:	romein	rill%ing	voet#bal	fone#tiek	dienst#plicht&ig	
morph 3:	romein	rill%ing	voet#bal	fonetiek	dienst#plicht&ig	
grafon:	ROOMEIN	RIL%IQ	VUT#BAL	FOONEETIIK	DIINST#PLIXT&CX	
stress 1:	ROOM'EIN	R'IL%IQ	V"UT#B"AL	FOONEET"IIK	D"IIINST#PL'IXT&CX	
stress 2:	ROOM'EIN	R'IL%IQ	V'UT#B"AL	FOONEET'IIK	D"IIINST#PL'IXT&CX	

REFERENCES

1. E. Berendsen, S. Langeweg, H. v. Leeuwen, The Proceedings of COLING 86 (Bonn, 1986), p. 613-617.
2. N. Chomsky, M. Halle, The Sound Pattern of English (New York, 1968).
3. R. Kager, E. Visch, W. Zonneveld, GLOT 10, (1987, to appear).
4. S. Langeweg, Linguistics in the Netherlands 1986 (Dordrecht, 1986), p. 151-161.
5. H. v. Leeuwen, S. Langeweg, E. Berendsen, The Proceedings of the IEE Conference on Speech Input/Output; techniques and applications (London, 1986), p. 200-206.