Rhythmic Factors in Weak-Syllable Insertion: An Internet Corpus Study

Hugo Quené

Utrecht inst of Linguistics OTS
Utrecht University, The Netherlands
hugo.quene@let.uu.nl

Abstract

Dutch language users often insert an inflectional schwa after an adverb, in certain grammatical constructions. The main hypothesis here is that this insertion, which is often ungrammatical, is driven by speakers' tendency towards regular speech rhythm, which overrides the fine grammatical nuances conveyed by absence of inflection. This rhythmicity hypothesis was investigated in a huge text corpus, viz. all web pages written in Dutch. The proportion of weak-syllable insertion was obtained for a sample of test phrases, varying in rhythmic context around the insertion point. Logistic regression of these proportions shows large and significant effects of rhythmic context on the odds of weak-syllable insertion. Hence, this insertion may well be due to rhythmical factors in language production, in addition to lexical-grammatical factors.

1. Introduction

Dutch language users often confuse the adverb *heel* /hel/ 'very, whole' and the inflected adjective *hele* /helə/ 'whole', in phrases such as *hele mooie jurken* (the normally intended gloss is 'very nice dresses', not 'whole nice dresses'). This particular violation of traditional grammar seems to be highly noticeable, and it is indeed often mentioned in language fora [12], in weblogs, by self-declared language purists, etc. The usual context is a phrase as the example above, crucially consisting of an adverb modifying a subsequent adjective, henceforth A–A phrases.

Dutch adverbs, unlike English ones, have no suffix *+ly*, and they are usually identical in form to adjectives. Adverbs are never inflected (according to conventional grammar), but adjectives can be inflected in Dutch. Hence, if the adjective is inflected and the preceding A is not, then the preceding A can only be an adverb modifying the adjective (e.g. *echt*+0 *hoog*+[ə] *bomen* 'really high trees'). If the adjective is inflected and the preceding A is also inflected, however, then that preceding A can only be an adjective modifying the noun (e.g. *echt*+[ə] *hoog*+[ə] *bomen* 'real high trees'); this is referred to as 'attributive' usage.

The conventional standard grammar for Dutch states that this latter attributive usage is acceptable only in informal language [3]. The adverbial usage, without inflectional schwa, is preferred in written language [12].

So far, however, one obvious factor governing this phenomenon has been overlooked, viz. speakers' tendency towards regular speech rhythm. In the word sequence *echt hoge bomen*, a "stress clash" occurs between the two A words. This stress clash could be resolved by schwa insertion, yielding an alternating sequence of strong and weak syllables. Presumably, such an alternating pattern is easier to pronounce, at least in English and Dutch.

Hence, the rhythmic tendency underlying schwa insertion may well be the same that underlies stress shift, according to "rhythmic" explanations of stress shift [4, 11, 1]. The main difference is that rhythmicity is not obtained by shifting the stress position in the first word (as in stress shift), but by inserting a weak inflection (schwa) syllable after that word, even though this may be grammatically inappropriate. Hence, the phenomenon is essentially one of weak-syllable insertion (or schwa epenthesis), and not one of adverb/adjective confusion.

The present study investigates the hypothesis that this weak-syllable insertion is due to speakers' tendency towards rhythmic alternation, even when the preceding A is intended as an adverb. (By analogy, this tendency would predict adverbial realizations like *real+ly* in English, even where adjectival *real*+0 would be intended, to achieve rhythmic alternation.) According to this hypothesis, this weak-syllable insertion in A–A phrases is similar to within-word epenthesis of a svarabhakti vowel (i.e., to schwa insertion) in Dutch, as in *film, melk*, which also depends on the rhythmic context of the insertion point [7]. The resulting sequence of strong and weak syllables is presumably easier to pronounce, and words with epenthesised schwa are also easier to perceive in the appropriate rhythmic context [2].

In fact, two variants have been postulated for the rhythmic tendency introduced above. The first, weaker tendency is towards *alternation* of strong and weak syllables [4, 7], without any restrictions on the actual timing in speech production. The second, stronger tendency is towards *isochrony*, i.e. towards equal inter-stress intervals in actual speech. Both accounts assume that speakers tend to place stresses at periodically spaced temporal locations. Testing the latter variant requires experimental studies. Indeed, some phonetic analyses of stress-shifted speech seem to support the stronger "isochrony" tendency [11, 1]. For the future, schwa insertion may well be investigated by means of phonetic analyses, of preferably spontaneous speech (because schwa insertion is regarded as ungrammatical in more formal speech [3]). The present study is more modest in scope, however, and attempts to provide preliminary evidence for the weaker "alternation" tendency, using written language only.

One might argue that conclusions about rhythmic factors cannot be derived validly from written language materials. If the written materials show rhythmic effects, however, then the most economical explanation would be to attribute such rhythmic effects to the underlying phonological attributes of the utterances written down in the corpus. Of course, these rhythmic effects, if any, should be further corroborated in speech corpora and by means of phonetic experiments.

Hence, the present study investigates the occurrence of schwa syllable insertion in written A–A phrases. The rates of occurrence are determined by means of a huge corpus, viz. all

web pages written in Dutch [9]. A major advantage of this corpus is the easy access to this material, and the huge size of the corpus. The search engine used in this study reported an estimated 225 million Dutch web pages containing any of the articles *de, het, een* 'the, the, a'; this provides us with a rough estimate of the size of the corpus. A disadvantage, however, is that the counts are not fully reliable. Repetitions of the same phrase in a single web page typically count as a single token (e.g. *...een hele strenge school met hele strenge regels...*). The search engine might also erroneously report multiple web pages containing the same text (e.g. minor updates) as multiple tokens. Third, the search engine itself may be unreliable, especially when its counts are extrapolated from a fraction of indexed web pages [6, 13]. Presumably, however, the huge size of the Internet corpus will outweigh these disadvantages.

2. Method

2.1. Sample phrases

The sample phrases to be investigated in the corpus were constructed from 17 adverbs (henceforth A1) that can modify the following adjective, and 13 following adjectives (henceforth A2) that are always inflected, yielding $17 \times 13 = 221$ sample phrases (see Table 1.) According to traditional grammar [3], inflection of A1 indicates attributive, adjectival use of A1, as opposed to uninflected, adverbial use. The probability of this attributive use depends on the meaning of the words in the sample phrase. A typical adverb like *tamelijk* 'fairly' cannot be used as an adjective, whereas such adjectival or attributive use is fairly likely for words like *bijzonder, enorm* 'particular(ly), enormous(ly)'. Hence, there are considerable lexical differences among phrases in their expected rates of occurrence of schwa insertion.

The stress pattern in the A1 and A2 words constitutes a critical factor in this study, because this is hypothesized to affect the rate of occurrence of schwa insertion. Hence, monomorphemic A1 words were chosen with stress on either the final, penultimate, or antepenultimate syllable (plus one compound A1, having ante-antepenultimate stress). Similarly, A2 words were chosen with stress on either the first, second, or third syllable. These stress positions were recoded as the number of syllables following the stressed syllable (for A1) and preceding the stressed syllable (for A2), because this was hypothesized to be the critical factor for schwa syllable insertion.

2.2. Phrase counts and analysis

Two variants of each test phrase, with and without insertion of inflection schwa, were input to the Yahoo! internet search engine (`www.yahoo.com`) in appropriate orthography. Detailed analyses of search results have indicated that Yahoo! counts are more reliable than those by Google [6, 13]. All searches were restricted to web pages written in Dutch, with target words input as phrases (two words in order). Searches were conducted on Nov 30 and Dec 1–2, 2005, from several computers in Utrecht, The Netherlands. The resulting counts (estimates reported in the Yahoo! search results) were noted down, for both variants of each test phrase. The counts were then converted to proportions of schwa insertion, for each phrase.

For 50 out of 221 test phrases, the observed "phrase frequency", i.e., the added counts of the two variants, turned out to be zero. These phrases were discarded from further analysis, because they provide no information about weak-syllable insertion. The remaining 171 phrases varied widely in phrase frequency, ranging from 1 (e.g. *matig voorzichtige* 'moderately careful') to 543000 (for phrase *heel* grote*). One could argue that phrases with low counts are poor test phrases, because they cannot inform us reliably about the occurrence of weak-syllable insertion. For this reason, phrases with low counts, ranging from 1 to 29 (n=79), were also discarded. This left only n=92 remaining test phrases with substantial Yahoo! counts for further analysis.

Table 1: *Selected Dutch words used for constructing sample phrases, with number of post-stress (A1) and pre-stress (A2) syllables.*

A1			A2
absurd	0	0	diepe
echt	0	0	grote
enorm	0	0	kleine
erg	0	0	rode
heel	0	0	strenge
intens	0	1	concrete
onwijs	0	1	gezonde
specifiek	0	1	speciale
bijzonder	1	1	urgente
gigantisch	1	1	verkeerde
matig	1	1	verstandige
ontzettend	1	1	voorzichtige
typisch	1	2	adequate
ongekend	2		
tamelijk	2		
vreselijk	2		
buitengewoon	3		

3. Results

The observed proportions of weak-syllable insertion, averaged over phrase types and broken down by the number of post-stress (*postS*, for A1) and pre-stress syllables (*preS*, for A2), are given in Table 2.

Table 2: *Proportions of weak-syllable insertion (with number of phrase types per cell, in parentheses), broken down by the number of post-stress syllables (*postS*, for A1) and pre-stress syllables (*preS*, for A2). Proportions are based on types of phrases ($n = 92$), not weighted by frequency (see text).*

	postS			
preS	0	1	2	3
0	.39 (32)	.39 (17)	.15 (12)	.11 (4)
1	.23 (20)	.26 (5)	(0)	(0)
2	.22 (2)	(0)	(0)	(0)

These results show that in general, the proportion of schwa insertion decreases as the first word ends in more unstressed syllables (differences between columns, in Table 2), and as the second word begins with more unstressed syllables (differences between rows). This decrease of weak-syllable insertion supports the main hypothesis, that speakers' tendency towards rhythmic alternation is a relevant factor in this insertion. If the first word results in more unstressed syllables, then the resulting syllable sequence conforms better to the preferred rhythmical alternating pattern. Consequently, there is less need for schwa insertion, and lower odds of schwa insertion are indeed observed.

In order to verify these apparent trends, the proportions were input into a logistic regression model [5, 10], with the "rhythmic" factors *postS* and *preS* and their interaction as predictors. The dependent variable in this analysis is the logit of the proportions of schwa insertion. The logit of proportion P is defined as the logarithm of the odds of P, or $\text{logit}(P) = \log(P/(1-P))$. Logistic regression is perfectly suitable for regression analysis with a dichotomous (binomial) dependent variable, as in this study. In addition, it allows the modelling of several factors in a single analysis, which is not possible in non-parametric analyses (which also happen to be less powerful). The output of this logistic regression consists of estimated regression coefficients for each linear predictor. Essentially, logistic regression attempts to model the logit data, given the independent variables, by estimating the best-fitting regression coefficients for their effects.

The resulting regression coefficients did not show any significant effects. The estimated intercept, for test phrases with postS $= 0$ and preS $= 0$ (e.g. *echt hoge*), yielded a value of -0.368 logit units (corresponding to the average proportion of .39 above, s.e. 0.336 logit units, $t = -1.09$, n.s.). Contrary to predictions, neither of the two rhythmic predictors, i.e. the number of post-stress syllables in A1 (*postS*), and of pre-stress syllables in A2 (*preS*), contributed significantly to the logistic regression (*postS*: $-0.499, t = -1.57$, n.s.; *preS*: $-0.763, t = 1.37$, n.s.). The apparent effects in Table 2 thus fail to reach significance, presumably due to the low numbers of phrase types.

Thus, one problem with these results lies in their low number of observations (phrase types). An additional problem is that the frequencies of usage vary widely among the 92 phrases, and that these frequency differences were ignored. One could well argue that this distorts the data, and that one should weight the proportion-of-insertion of each phrase according to the frequency of that phrase. The adjusted proportions of weak-syllable insertion, again broken down by the rhythmic factors *postS* and *preS*, are given in Table 3. These proportions are now based on *tokens* of test phrases, i.e., weighted for phrase frequency.

Table 3: *Proportions of weak-syllable insertion (with number of tokens per cell, in parentheses), broken down by the number of post-stress syllables (*postS*, for A1) and pre-stress syllables (*preS*, for A2). Proportions are based on tokens of phrases (estimated $N = 1184664$).*

	postS			
preS	0	1	2	3
0	.64	.09	.07	.06
	(1050391)	(24278)	(5312)	(1532)
1	.42	.41		
	(102634)	(436)	(0)	(0)
2	.20			
	(81)	(0)	(0)	(0)

These frequency-weighted (token) proportions of weak-syllable insertion show far stronger effects of rhythmic context, as compared to the unweighted (type) proportions in Table 2. The proportion of weak-syllable insertion clearly decreases with the number of unstressed syllables in the preceding word (*postS*, columns) and in the following word (*preS*, rows). Insertion occurs most often between two adjacent stressed syllables. If there is less need for schwa insertion, because there are more unstressed syllables around the A–A word boundary, then lower odds of a schwa being inserted are indeed observed. Thus the observed proportions provide supporting evidence for the main hypothesis of this study, viz. that rhythmic factors contribute to the odds of weak-syllable insertion.

For high-frequency phrases (with many tokens in Table 3), their token-based proportions of insertion are *higher* than the former type-based proportion (Table 2); for low-frequency phrases the proportions are *lower*. This suggests that phrases with higher frequency are more prone to weak-syllable insertion. This may not be a coincidence, but a natural effect of phrase frequency, if speakers insert weak syllables to get a better rhythm in their speech. (The effect of phrase frequency on the proportion-of-insertion was investigated for each phrase *type*, using logistic regression. Neither phrase frequency nor the rhythmic factors showed significant effects, again because of the low number of phrase types. A larger sample of test phrases, with sufficient variation in phrase frequency, is needed to assess the effect of frequency on weak-syllable insertion.)

The frequency-weighted (token) proportions were fed into a logistic regression model, with the rhythmic factors *postS* and *preS* and their interaction as predictors. The resulting regression coefficients (in logit units) are reported in Table 4.

Table 4: *Logistic regression coefficients (with standard error and* t *test statistic), for the number of post-stress syllables (*postS*, for A1), pre-stress syllables (*preS*, for A2), and their interaction. Regression was based on frequency-weighted (token) proportions of schwa insertion, for 92 remaining test phrases (see text).*

predictor	coef.	(s.e.)	*t*
(Intercept)	0.572	0.002	281.7
postS	-2.499	0.018	-137.4
preS	-0.891	0.007	-134.3
postS × preS	2.446	0.099	24.6

Both rhythmic factors of interest, *postS* and *preS*, yield highly significant coefficients in this second regression analysis, with $p < .0001$ for both. This confirms the strong effects of rhythmic context on weak-syllable insertion (Table 3).

The negative coefficients also show that the effect of A1 stress position (*postS*) is larger than that of A2 (*preS*). Because the inserted schwa is attached to A1 and not A2, it is appropriate that the stress pattern of the so-inflected A1 has a larger influence than that of the following word. However, the smaller effect size for predictor *preS* could also be due to sampling fluctuation in the actual A2 words chosen. In particular, only one A2 word was selected that has *two* pre-stress syllables (i.e., having stress in its third syllable, see Table 1). The smaller range of *preS* values, and lower number of tokens, may have limited the effect of this predictor.

Finally, the significant interaction effect suggests that the odds of schwa insertion are higher, if there are both more post-stress syllables (A1) and more pre-stress syllables (A2). This interaction is due to the cell with both postS $= 1$ and preS $= 1$, where the proportion-of-insertion of .41 is higher than expected. Closer inspection of these cases revealed one test phrase with particularly high odds of insertion, viz. *bijzonder* speciale* 'particular(ly) special', which accounted for over half of the phrase tokens in this cell. In this particular phrase, the first A1 may well have been *intended* to be inflected, for semantic reasons, i.e. to indicate attributive usage of A1. This is certainly the intended use in one idiomatic expression containing

this phrase, viz. *bijzondere speciale scholen* 'denominational special schools'. This single idiosyncracy may well explain the unusually high odds of attributive, inflected use of A1 in this cell.

In order to verify this explanation, the proportions were re-analyzed after this single phrase was discarded. The observed proportion-of-insertion for that cell then becomes 22% (over $n = 211$ tokens), which is closer to the expected proportion without interaction. The frequency-weighted logistic regression was also re-run without this single test phrase. The resulting intercept and main effects were virtually similar to those given in Table 4. The interaction effect, however, yielded a considerably smaller regression coefficient than before (1.568, s.e. 0.167), although still significant ($t = 9.4, p < .0001$). This considerable decrease in proportion and in regression coefficient suggests that the unexpected interaction effect may indeed have been due to lexical and idiomatic idiosyncracies in the test phrases, and that it may disappear if a larger sample of test phrases will be investigated.

4. Discussion and Conclusion

The results suggest that weak-syllable insertion is indeed sensitive to speech rhythm. The odds of weak-syllable insertion are highest if there is a "stress clash" at the insertion point, and the odds decrease if there are more unstressed syllables at the insertion point. Hence, these results are in agreement with the rhythmic-tendency hypothesis of this study. In conclusion, weak-syllable insertion seems to be due, at least in part, to speakers' tendency towards rhythmic speech.

One might argue that the present study does not allow conclusions about rhythmical tendencies in speech behavior, because it involves analysis of written corpus materials only. It is certainly true that one should be cautious in generalizing from written to spoken language. In this study, however, such caution is already built in the corpus material of written language. Because written language is intrinsically more formal than spoken language, a lower incidence of weak-syllable insertion may be expected in the presently used Internet corpus, as compared to a spoken language corpus [3]. The observed ubiquity of this insertion in written language indicates that the grammatical distinction (as outlined in the introduction) may well be lost to many Dutch language users. Hence, the results from this corpus of written Dutch may also generalize to spoken Dutch. The present results of this study need to be validated with the Corpus of Spoken Dutch [8], of course, although the far smaller size of the latter corpus (by many orders of magnitude) is likely to hamper statistical analysis.

The text-based results presented above are in agreement with both the "alternation" and "isochrony" accounts of the rhythmic-tendency hypothesis. Data about actual speech timing are needed to distinguish these two accounts. Obviously, this requires controlled experiments, like those demonstrating similar rhythmicity effects in stress shift [11].

In summary, the present study shows that weak-syllable insertion in Dutch A–A phrases is influenced by the rhythmic context at the insertion point. This in turn suggests that speech rhythm is an important factor here, in addition to lexical-semantic and grammatical factors.

5. Acknowledgements

My thanks are due to Frank Jansen, Sieb Nooteboom and Esther Janse, for helpful comments and suggestions.

6. References

[1] Barbosa, P.A.; Antares, P.; Silveira, L.S., 2004. Unifying stress shift and secondary stress phenomena with a dynamical systems rhythm rule. Paper presented at the 2nd Int Conf on Speech Prosody, 23-26 March, Nara, Japan.

[2] van Donselaar, W.; Kuijpers, C.; Cutler, A., 1999. Facilitatory effects of vowel epenthesis on word processing in Dutch. *Journal of Memory and Language*, 41(1), 59-77.

[3] Haeseryn, W.; Romijn, K.; Geerts, G.; de Rooij, J.; van den Toorn, M.C., eds., 1997. *Algemene Nederlandse Spraakkunst* (2nd rev.ed.). Groningen: Nijhoff. [http://www.ru.nl/e-ans]

[4] Hayes, B., 1984. The phonology of rhythm in English. *Linguistic Inquiry*, 15, 33-74.

[5] Hosmer, D.W.; Lemeshow, S., 2000. *Applied Logistic Regression*. New York: Wiley.

[6] Liberman, M., 2005. More arithmetic problems at Google. *Language Log*, 1840. http://itre.cis.upenn.edu/~myl/languagelog/archives/001840.html.

[7] Kuijpers, C.; van Donselaar, W., 1998. The influence of rhythmic context on schwa epenthesis and schwa deletion in Dutch. *Language and Speech*, 41, 87-108.

[8] Oostdijk, N., (2000). Het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 5(3), 280-284.

[9] van Oostendorp, M.; van der Wouden, T., 1998. Corpus Internet. *Nederlandse Taalkunde*, 3(4), 347-361. [http://www.niederlandistik.fu-berlin.de/digitaal/digitaal-04.html]

[10] Pampel, F.C., 2000. *Logistic regression: A primer*. Quantitative applications in the social sciences; 07-132. Thousand Oaks, CA: Sage.

[11] Quené, H.; Port, R.F., 2002. Rhythmical factors in stress shift. In *Papers from the 38th Meeting of the Chicago Linguistic Society (Volume 1: The Main Session)*, M. Andronis, E. Debenport, A. Pycha, K. Yoshimura (eds.). Chicago: Chicago Linguistic Society, 549-562. [http://www.indiana.edu/~iulcwp, #02-18A].

[12] Taaladviesdienst, s.a. Heel/hele mooie jurken? *Onze Taal*. [http://www.onzetaal.nl/advies/heel.htm]

[13] Véronis, J., 2005. Web: Google's counts faked? [http://aixtal.blogspot.com/2005/01/web-googles-counts-faked.html]