Segment Deletion in Spontaneous Speech: A Corpus Study using Mixed Effects Models with Crossed Random Effects

Christophe Van Bael [1], *Harald Baayen* [2], *Helmer Strik*[1]

[1] Centre for Language and Speech Technology, Radboud University Nijmegen, the Netherlands
[2] Interfaculty Research Unit for Language and Speech, Radboud University Nijmegen, the Netherlands

c.v.bael@gmail.com, baayen@mpi.nl, strik@let.ru.nl

10.21437/Interspeech.2007-713

Abstract

We studied the frequencies of phone and syllable deletions in spontaneous Dutch, and the extent to which such deletions are influenced by the various linguistic and sociolinguistic factors represented in the transcriptions, word segmentations and metadata of the Spoken Dutch Corpus. In addition to providing insight into the frequencies of phone and syllable deletions and the factors influencing them, our study illustrates the new opportunities for analysing rich and therefore complex corpus data offered by a recently developed statistical modelling technique: the possibility to model the effects of random factors as crossed instead of nested with generalised linear mixed effects models.

We observed average phone and syllable deletion rates of 7.57% and 5.46% respectively. 20.32% of the words had at least one phone missing, and 6.89% of the words had at least one syllable deleted. The mixed effects models for phone and syllable deletion had several effects in common, which implies that both types of deletion are to a large extent influenced by the same factors. The strongest factors across both models were lexical stress, word duration and the segmental context of the syllable onset of the following word.

Index Terms: segment deletion, corpus linguistics, statistical modelling.

1. Introduction

Over the years, large phonetically transcribed speech corpora have proven valuable resources for studying pronunciation variation. Switchboard [1] and the Buckeye Corpus of Conversational Speech [2], to name just two examples, have proven useful for -among other things- creating an inventory of testified speech processes in everyday conversational English [1], studying the frequencies of these processes [3] and investigating how these processes are influenced by various linguistic and socio-linguistic factors (e.g. [4],[5]). Because most phonetically transcribed speech corpora comprise (American) English, most corpus studies on pronunciation variation were conducted on English. The recent release of the richly annotated 9-million-word Spoken Dutch Corpus [6] (CGN) now offers new opportunities for studying pronunciation variation in a language other than English, and for testing whether knowledge gleaned for American English also holds for another language.

The first aim of our study was to establish the frequencies of segment deletions in spontaneous Dutch, and the extent to which such deletions are influenced by the linguistic and sociolinguistic factors reflected in the annotations, word segmentations and metadata of the CGN. We defined segment deletion as the deletion of phones and syllables that can be inferred from the symbolic alignment of canonical and manually verified phonetic transcriptions from the so-called *core corpus* of the CGN.

An ancillary goal of our study was to explore the new opportunities for analysing complex corpus data offered by a recently developed statistical modelling technique: the possibility to model the effects of random factors as crossed instead of nested with generalised linear mixed effects models (GLMMs) [7]. Mixed effects models are interesting for linguistic corpus studies because they allow for the inclusion of factors with repeatable levels (e.g. word class) and randomly sampled levels (e.g. speaker) in the same model, because they can cope with missing data and with complex factorial designs, and because they can do all this in a computationally efficient way [8]. Until recently, however, factors with randomly sampled levels could only be modelled with nested designs. This imposed serious limitations on the use of mixed-effects models for linguistic studies because it could result in anti-conservative P-values for the fixed effects in the models. This could lead to fixed effects being declared statistically significant too easily in addition to the random effects in the models. In other words, it could increase the risk of type I errors, i.e. erroneously considering an effect significant. The recent possibility to model random effects as crossed instead of nested alleviates this problem [7].

This paper is organised as follows. Section 2 presents our methodology. In Section 3, we present and discuss the results of our analyses. Section 4 summarises our conclusions.

2. Methodology

2.1. Data preparation

We based our study on the annotations, word segmentations and metadata of spontaneous telephone dialogues in the CGN. Excluding broken and (partially) unintelligible words, we obtained a dataset of 178,271 word tokens (8,539 types) with manually verified word boundaries, orthographic and broad phonetic transcriptions and POS tags. Similar to [3], we generated a canonical representation of the material by concatenating the citation forms of the words. These citation forms (including syllable boundaries and lexical stress marks) were retrieved through lexicon lookup. We identified phone deletions by aligning the phones in the canonical and manually verified phonetic transcriptions with ADAPT [9]. In the same process, the syllable boundaries (and lexical stress marks) of the canonical representation were copied onto the manually verified transcriptions to identify syllable deletions (see Figure 1).

CT	Ei	G @	l @	k
PT	Ei	- -	l @	k

Figure 1: *Identification of two phone deletions and one syllable deletion through the alignment of a canonical transcription (CT) and a broad phonetic transcription (PT).*

For every canonical phone and syllable, we derived information at the utterance, word, syllable and phone level. In addition, sociolinguistic (speaker) information was extracted from the metadata. All this information was stored in a separate *information vector* for every canonical phone and syllable.

At the utterance level, we considered the duration in ms (excluding silent pauses) and the number of canonical phones and syllables. From this information we computed the articulation rate in phones and syllables per second. At the word level, we considered the word identity, the word duration, the number of canonical phones and syllables, the position in the utterance (initial, final, initial-final, mid), the word class (nouns, verbs, adjectives, adverbs, pronouns, interjections, articles, numerals, conjunctions, prepositions), and the number of times the word was previously uttered by one of the interlocutors and by the current speaker (to model the effects of given/new information). We also considered the word's frequency and the mutual information of the word and its neighbours (both computed on the orthographic transcription of the 544,215 word tokens in the telephone dialogues of the CGN that were *not* included in the core corpus), whether the word preceded a long silence (>250 ms) or a disfluency (repetition, broken word, filled pause) and whether the following word started with a consonant or a vowel. At the syllable level, we considered the syllable identity, the syllable's position in the word (initial, final, initial-final, mid), the number of canonical phones, and whether the syllable had lexical stress (retrieved through lexicon lookup). At the phone level, we considered the phone identity, its position in the word (initial, final, initial-final, mid), syllable (initial, final, initial-final, mid) and in the consonant/vowel structure of the syllable (e.g. CC_V), whether the phone was part of the syllable's onset, nucleus or coda and whether the phone had lexical stress (retrieved through lexicon lookup). In addition, we considered the identity of the speaker, his or her gender, age (year of birth), regional background (the region the speaker spent most of the time between the age of 4 and 16) and level of education (high, mid, low). In the last field of each information vector, we marked whether the phone or syllable was deleted. Like [3], we considered syllables deleted if their syllabic nucleus was absent. Contrary to English, Dutch normally does not have syllabic nasals, laterals or rhotics. Therefore we considered syllables deleted if a vocalic nucleus was no longer present.

2.2. Analyses

We first counted the number of phone and syllable deletions. The results of this analysis are presented in Section 3.1. Subsequently, we fitted two GLMMs with a logistic link function to the information vectors: a model for phone deletion and a model for syllable deletion. We assumed binomial variance. Both models were defined by sequentially including every linguistic and sociolinguistic factor from the information vectors in the model. A factor was only retained if it contributed significantly ($p < .05$) to the model's goodness of fit. Factors were pruned from the model if their contribution was no longer significant after the inclusion of additional factors. Goodness of fit was assessed with Somers' Dxy, a rank correlation between predicted probabilities and observed responses which is closely related to the receiver operating characteristic (ROC) curve area [10]. The results of the statistical analyses are presented in Section 3.2. All statistical computations were conducted with the `lme4` package for R [11]. We fitted models on a randomly selected 10% subset of the material to keep model building computationally feasible.

3. Results

3.1. Frequencies of segment deletions

We counted 42,556 phone deletions out of 562,294 phones in 85,050 content and 93,221 function words. This implies an overall phone deletion rate of 7.57%. 83.69% of all phone deletions concerned deletions of one of the 7 following phones: /@/ (22.91% of all deletions), /r/ (19.59%), /n/ (13.12%), /t/ (12.27%), /l/ (5.74%), /h/ (5.39%) and /d/ (4.66%). This was not just because these phones are common in Dutch; the relative deletion rates of these phones proved higher than the deletion rates of other phones. We found the following proportion of deletions for each of these phones: /r/ (28.79% - 95% confidence interval: 1.05), /h/ (21.25% - 1.55), /@/ (16.10% - 0.59), /n/ (12.40% - 0.61), /l/ (12.29% - 0.92), /t/ (11.40% - 0.58), /d/ (7.06% - 0.60).

When assessing phone deletion at the word level, we found that 17.56% of the words had one phone missing, 2.16% had two phones missing, and 0.60% had three or more phones missing. We found relatively more phone deletions in function words (8.59%) than in content words (6.49%), but at the same time we observed more individual content words than function words with phone deletions (22.76% vs. 18.10%). This implies that the phone deletions in the function words were concentrated in a proportionally smaller subset of the words than was the case with the content words. Just like [3], we have to conclude that, given the short average word length (in our study 3.14 canonical phones per word) the proportion of words with one or more missing phones is remarkably large. Table 1 illustrates how the phone deletions were distributed in words of different length.

Table 1. *Frequencies of phone deletion. 95% confidence intervals between brackets.*

	content words		function words	
# syl/word	# phones	% deleted	# phones	% deleted
1	155,463	6.82 (0.25)	191,208	8.66 (0.25)
2	108,329	6.02 (0.28)	18,813	8.39 (0.80)
3	61,444	9.27 (0.46)	2,113	5.35 (1.97)
4	17,515	5.85 (0.70)	1,351	6.88 (2.78)
5 or more	5,568	5.85 (1.25)	490	6.73 (4.67)
# phn/word	# phones	% deleted	# phones	% deleted
1	24	0.00	10,080	6.14 (0.95)
2	23,520	3.21 (0.46)	111,294	7.24 (0.31)
3	101,862	7.74 (0.33)	66,978	11.56 (0.49)
4	52,624	6.13 (0.41)	11,492	5.89 (0.87)
5	46,035	5.78 (0.43)	5,220	4.96 (1.20)
6	42,222	8.20 (0.53)	4,284	8.52 (1.70)
7	27,377	7.23 (0.62)	2,191	22.23 (3.53)
8	22,664	8.81 (0.74)	280	7.86 (6.72)
9	13,860	7.61 (0.89)	180	1.67 (4.76)
10 or more	18,131	6.38 (0.72)	1,976	7.79 (2.42)
total	348,319	6.94 (0.17)	213,975	8.59 (0.24)

Table 1 shows both for content and function words decreasing phone deletion rates in monosyllabic and bisyllabic words and words with 4 syllables or more. Contrary to the general trend of function words being more prone to deletions than content words, trisyllabic content words were remarkably more susceptible to phone deletion than trisyllabic function words. Moreover, trisyllabic content words were more susceptible to phone deletion than other content words, whereas trisyllabic function words had relatively fewer phones deleted than the other function words. Common trisyllabic adverbs such as 'allemaal', 'helemaal', 'inderdaad' and adjectives such as 'natuurlijk', 'eigenlijk',

'allerlei' were particularly susceptible to phone deletion. Content words with two phones or less were less susceptible to phone deletion than content words with three or more phones. This can probably to a large extent be explained by the frequent use of discourse words like 'ja' (yes) and 'mm' (uhu), which often constitute a speech utterance of their own and are therefore well pronounced. We noticed increasing deletion rates for 1, 2, and 3-phone function words, with a remarkably high number of phone deletions in 3-phone function words (11.56%). This high deletion rate was largely due to the frequent occurrence of final /r/-deletions in common words such as 'naar' (towards), 'daar' (there), 'voor' (for) and due to the frequent use of words such as 'm'n' (from: 'mijn' - my) and 'z'n' (from 'zijn' - his) in which the canonical nucleus /@/ was deleted. These observations largely explain the high average phone deletion rates of /r/ (28.70%) and /@/ (16.10%) reported before and the high average syllable deletion rates in monosyllabic function in Table 2.

Inspection of the average phone deletion rates per word class proved articles, pronouns and conjunctions most prone to deletion. This is in line with the findings of [12] for American English. Interjections, nouns and numerals were least susceptible to phone deletion. The low deletion rate of the interjections can be largely explained by the frequent use of filled pauses, which by default were transcribed by means of their citation form /@/. The low deletion rate of both nouns and numerals can be explained by the high information valence associated with these words.

12,534 out of 229,670 syllables (5.46%) were deleted. As with the frequencies for phone deletion, we observed relatively more syllable deletions in function words (6.68%) than in content words (4.54%). 7.09% of the function words had 1 syllable missing, and a negligible 0.01% had 2 syllables missing. 6.33% of all content words was pronounced with 1 syllable missing, 0.33% had 2 or more syllables missing. The relatively high syllable deletion rate in function words in our data can be explained by the frequent use in Dutch of contracted words such as 'm'n' and 'z'n' in which the canonical syllabic nucleus /@/ was often deleted. These deletions accounted for the deletion of 2.62% of the syllables in the function words. The remaining 4.06% (6.68% – 2.62%) syllable deletion rate is comparable to the 4.5% syllable deletion rate reported for function words in American English [3].

Table 2 shows the distribution of syllable deletions over N-syllable content and function words. Not surprisingly, the deletion of just one syllable was more common than the deletion of more syllables. Only 1.65% of the syllabic nuclei in the monosyllabic content words were deleted whereas 6.98% of the syllabic nuclei in monosyllable function words were deleted.

3.2. Modelling segment deletion

We first fitted a GLMM to the data with speaker, phone, syllable, word, syllabic structure and regional background of the speaker as crossed random effects and phone deletion as response variable. Somers' Dxy of the final model was equal to 0.86 (ROC curve area = 0.93) which indicates that the model provided a good fit to the data. Most fixed-effects predictors were significant at at least the 0.01 level. Inclusion of the random effects in the model was supported by likelihood ratio tests (ANOVA tests, all p-values < 0.05). In addition to the effects of the phone identity ($\hat{\sigma}$ = estimated standard deviation of the random effect = 1.56), syllabic structure ($\hat{\sigma}$ = 1.06) and the regional background of the speaker ($\hat{\sigma}$ = 0.16) which were treated as random-effects factors to limit the number of parameter estimates in the model, we observed main effects for eight fixed-effects factors over and above the random variation that came with the speakers ($\hat{\sigma}$ = 0.24) and items (words ($\hat{\sigma}$ = 0.86), syllables ($\hat{\sigma}$ = 0.93)) sampled.

Table 2. *Frequencies of syllable deletion. 95% confidence intervals between brackets.*

content words				
# can syl	totals	no del(%)	1 syl del(%)	≥2 syl del(%)
1	52,638	98.35 (0.22)	1.65 (0.22)	
2	43,460	94.34 (0.44)	5.66 (0.44)	
3	25,224	91.80 (0.68)	6.42 (0.61)	1.78 (0.33)
4	7,324	94.51 (1.06)	4.61 (0.98)	0.86 (0.43)
5	1,795	93.70 (2.31)	4.35 (1.95)	1.95 (1.29)
6	360	95.00 (4.85)	3.89 (4.37)	1.11 (2.66)
7	105	98.10 (7.05)	1.90 (7.05)	
8	56	98.21(10.72)	1.79(10.72)	
9	18	100		
function words				
# can syl	totals	no del(%)	1 syl del(%)	2 syl del(%)
1	88,427	93.02 (0.34)	6.98 (0.34)	
2	8,632	95.69 (0.87)	4.19 (0.86)	0.12 (0.16)
3	963	96.47 (2.46)	3.12 (2.33)	0.42 (1.01)
4	484	97.52 (3.06)	2.48 (3.06)	
5	160	96.88 (6.37)	3.13 (6.37)	
6	24	100		

We observed main effects of word class, lexical stress (phones in stressed syllables were less likely deleted), the position of the phone in the syllable (greater likelihood of phone deletions in coda positions), the position of the phone in the word (deletions further into the word were more likely), the segmental context of the following word (words starting with consonants were more likely to induce phone deletion than words with vowels), the number of canonical phones in the word (words with more canonical phones were more likely to have phones deleted) and in the utterance (negative slope), and the duration of the word (the longer the actual duration of a word, the smaller the chance that phones were deleted).

Inspection of the best linear unbiased predictors (BLUPs) for the phone random effect (i.e. the by-phone adjustments to the overall intercept) revealed that, according to our model, /h/, /d/ and /@/ were most likely to be deleted, and that /p/, /k/ and /N/ were least easily deleted. The ordering of the BLUPs from highest likelihood for phone deletion to lowest likelihood for deletion generally agreed with the descending ordering of the individual deletion rates of the phones in our material. Inspection of the random intercepts for syllabic structure showed that phones were most likely to be deleted in C_CV, _C, CVC_C, and CVCC_ structures and least likely to be deleted in CV_C, CCV_C and V_ structures. Most of these findings were confirmed by our frequency counts. The susceptibility of the vocalic nucleus to deletion in VC (or: _C) syllables, which is clearly deviant from the findings of [12] for American English, can largely be explained by the frequent use of contracted words in Dutch. The deletion of schwa in such words accounted for 58.44% of all deletions in VC syllables.

The generalised linear mixed effects model for syllable deletion included speaker ($\hat{\sigma}$ = 0.29), word ($\hat{\sigma}$ = 1.73) and syllable ($\hat{\sigma}$ = 2.00) as crossed random effects and syllable deletion as response variable. Somers' Dxy was equal to 0.92 and the ROC curve area was equal to 0.96. These results indicate that also the syllable model fitted the data very well. All fixed-effects predictors were significant at at least the 0.01 level. Similar to our model for phone deletion, we observed main effects of word class,

lexical stress and word duration. Again syllabic stress rendered syllable deletion less likely, and longer word durations were strong cues for the preservation of syllabic nuclei. As opposed to what we saw in the model for phone deletion, words starting with a vowel increased the likelihood of syllable deletion in the previous word. Considering the high deletion rate of schwas, it is not unlikely that many schwas in unstressed and unaccented word-final syllables were deleted to ease the articulatory transition to the vowel of the next word. We also found that syllables were more likely deleted in utterances with more canonical syllables. Somewhat related, we noticed that a higher articulation rate rendered syllable deletions more likely. We also found that syllables in utterance-initial words were more prone to deletion than deletions in other words further down the sentence.

Because we could include random-effects factors such as speaker, word and syllable identity as crossed instead of nested (the same indirectly also holds for all the fixed-effects factors related to these random-effects factors), we were able to assess in a methodologically sound way the relative effect of every linguistic and socio-linguistic factor in the annotations, word segmentations and metadata of the CGN over and above the random variation that came with the speakers, words and syllables we sampled. In our study, it was interesting to analyse which factors were significant in the models, but it was equally interesting to see that the (potential) effects of factors which were previously reported to influence segment deletion were 'covered' by other factors. For example, mutual information (word predictability) which was previously reported to influence phone deletions (e.g. [4]) did not appear in our final model definitions, and word frequency was only significant in the phone deletion model. This may be due to several reasons. For example, we computed word frequency and mutual information on 'only' 544,215 words, and we chose to keep our models easily computable by not including correlations between factors. Computing estimates for word frequency and mutual information on a larger dataset and including correlations in the models may eventually render factors like word frequency and mutual information significant. We leave these issues open for further research. In any case, the absence of effects of e.g. word frequency and mutual information does not mean that these factors do not affect phone and syllable deletion. Rather, it implies that in our model definitions other factors showed a stronger effect on the deletion of phones and syllables. Actually, word frequency was part of the syllable model definition until we included 'word identity' as random effects factor. In both models, the effects of mutual information were probably covered by word frequency. Such knowledge is unlikely to be gained in controlled experiments on selected data sets aimed at studying the effects of one or a few factors at a time, but it can be of interest for pronunciation variation modelling of everyday conversational speech.

4. Conclusions

We studied the frequencies of phone and syllable deletions in spontaneous Dutch, and the extent to which such deletions are influenced by the interplay of linguistic and sociolinguistic factors than can be retrieved from the word segmentations, annotations and metadata in the Spoken Dutch Corpus. We found average phone and syllable deletion rates of 7.57% and 5.46% respectively. 22.76% of the content words and 18.10% of the function words had at least one phone missing, and 6.66% of the content words and 7.10% of the function words had at least one syllable missing. Even though these figures are lower than the figures reported in [3] for American English, our analyses just as well suggest that phone and syllable deletions are common in everyday conversational Dutch. The mixed effects models for phone and syllable deletion had several effects in common, which implies that both types of deletion are to a large extent influenced by the same factors.

Our study illustrates new opportunities for analysing rich corpus data by means of generalised linear mixed effects models with crossed random effects. The use of such statistical models is useful for exploratory research like ours as well as for hypothesis testing. The recent possibility to model random effects such as speaker and item in a principled way as crossed instead of nested makes it possible to ascertain in one model whether linguistic and sociolinguistic factors are predictive over and above the random variation that comes with the subjects and items sampled. As a consequence, linguistic phenomena such as segment deletion can now be studied in a methodologically sound way in corpus data as a function of the interplay of many factors instead of in controlled experimental environments designed for studying the effects of one or a few factors at a time.

5. Acknowledgement

The work of Christophe Van Bael was funded by the Speech Technology Foundation, Utrecht, the Netherlands.

6. References

[1] Greenberg, S., Hollenback, J., Ellis, D. "Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus", Proc. ICSLP, Philadelphia, USA, pp. S24-27, 1996.

[2] Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., Raymond, W. "The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability", Speech Comm., vol. 45/1, pp. 89-95, 2005.

[3] Johnson, K. "Massive Reduction in Conversational American English". Yoneyama, K., Maekawa, K. (Eds.) Spontaneous Speech: Data and Analysis. Tokyo: The National Institute for Japanese Language, pp. 29-45, 2004.

[4] Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D. "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation", JASA., vol. 113(2), pp. 1001-1024, 2003.

[5] Raymond, W.D., Dautricourt, R., Hume, E. "Word-internal /t,d/ deletion in spontaneous speech: Modeling the effects of extra-linguistic, lexical, and phonological factors". Language Variation and Change, vol. 18, pp. 55-97, 2006.

[6] Oostdijk, N. "The Design of the Spoken Dutch Corpus". Peters, P., Collins, P., Smith, A. (Eds.) New Frontiers of Corpus Research. Rodopi, Amsterdam, pp. 105-112, 2002.

[7] Baayen, R.H., Davidson, D.J., Bates, D.M. "Mixed effects modeling with crossed random effects for subjects and items", submitted manuscript, 2007.

[8] Pinheiro, J.C., Bates, D.M. "Mixed-effects models in S and S-PLUS". Statistics and Computing. Springer, New York, 2000.

[9] Elffers, B, Van Bael, C., Strik, H. "ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions". Internal report, Department of Language and Speech, Radboud University Nijmegen, the Netherlands, 2005.

[10] Harrell F.E. Jr, Lee K.L., Mark D.B. "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". Statistics in Medicine, vol. 15/4, pp.361–387, 1996.

[11] Bates, D.M., Sarkar, D. "lme4: Linear mixed effects models using S4 classes, R package version 0.9975-7", 2005.

[12] Greenberg, S. "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation". Proc. of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Kerkrade (Netherlands), pp. 47-56, 1998.