Assimilation of word-final nasals to following word-initial place of articulation in UK English

Margaret E. L. Renwick, Ladan Baghai-Ravary, Rosalind Temple, John S. Coleman

Phonetics Laboratory, University of Oxford, Oxford, United Kingdom

{margaret.renwick, ladan.baghai-ravary, john.coleman}@phon.ox.ac.uk

rosalind.temple@mod-langs.ox.ac.uk

Abstract

Using very large speech corpora, we can study rare but systematic pronunciation patterns in spontaneous speech. Previous studies have established that word-final alveolar consonants in English (/t/, /d/, /n/, /s/ and /z/) vary their place of articulation to match a following word-initial consonant, e.g. "ran quickly" → "ra[ŋ] quickly". Assimilation of bilabial or velar nasals, e.g. "alar[ŋ] clock" for "alarm clock", is unexpected according to linguistic frameworks such as underspecification theory. The existence of systematic counterexamples would challenge that theory, but these might have been previously overlooked because they are infrequent. From the c. 8-million word Audio BNC (http://www.phon.ox.ac.uk/AudioBNC) we extracted over 14,000 tokens of relevant word pairs, to determine whether non-alveolar assimilations occur and with what distribution. Word and segment boundaries were obtained by forced alignment, and F1-F3 formant frequencies were estimated using Praat. Formant frequencies in assimilation environments were compared to non-assimilating controls (e.g. *come back* vs. *come down*). We present evidence that velar and bilabial nasals sometimes do assimilate, though possibly less frequently than alveolars.

Index Terms: assimilation, nasal consonants, place of articulation, corpus

1. Introduction

The increasing number and size of speech corpora and advances in speech technology now provide unprecedented opportunities to study large quantities of real-life speech, and to answer linguistic questions which it has not been possible to address until now. More specifically, large corpora allow the investigation of phenomena which are systematic, and may therefore be relevant for modeling phonological processes, but which are also rare, and thus have previously lacked adequate empirical investigation. One such phenomenon, the object of this study, is the assimilation of word-final nasals to the place of articulation of following word-initial consonants.

According to many handbooks and textbooks on English phonology (e.g. [1], [2], [3], [4], [5], [6]), word-final alveolar consonants (i.e. /t/, /d/, /n/, /s/ and /z/) – and *only* alveolar consonants – vary their place of articulation to match the consonant with which the next word begins, e.g.:

that case	→	tha[k] case
ran quickly	→	ra[ŋ] quickly (*cf.* rang quickly)
this shop	→	thi[ʃ] shop (*cf.* fish shop)

Some phonological theorists have tried to explain the readiness of alveolar consonants to assimilate (vs. the resistance of velar and labial articulations to assimilation) by proposing that alveolar consonants are *underspecified* for (i.e. they *lack*) place of articulation features ([7]). Under this account, assimilation of labial or velar consonants should not happen; pronunciations such as "ki[m]pin" for "kingpin" or "alar[ŋ] clock" for "alarm clock" would be counterexamples. Assimilation of word-final alveolars to following consonants has been studied in an American English corpus by [8], but that paper did not look for instances of bilabial or velar assimilation. In this paper we show that assimilation of word-final labial and velar nasal consonants does in fact occur, albeit not very commonly, so that it has been largely overlooked in the literature. There is some prior evidence that velar and labial final consonants sometimes assimilate in English, e.g. I'm going → I'[ŋ] going ([9]), some girls → so[ŋ]e girls ([10]). However, these are anecdotal observations, with no context, no audio available for detailed study, and no statistics on their frequency of occurrence. In our current project, therefore, we systematically search for and assess the occurrence of such non-alveolar assimilations in a large corpus of natural speech, focusing on the contexts in which they occur most readily, and their acoustic characteristics.

2. The Audio British National Corpus

The Audio BNC uniquely provides the necessary resources for studies such as ours, in two ways: its relative informality and its size. Non-canonical assimilations may be far less likely to occur in careful, laboratory speech. Large size is needed because of the rarity of non-canonical assimilation, and because of the extremely unbalanced distributions of linguistic units (phonemes, syntactic constructions, words) in natural language — by Zipf's Law, some sounds, words, pairs of words, etc. are vastly more frequent than others. Containing over 1200 hours of recorded speech, the Audio BNC is the largest snapshot of transcribed "language in the wild" ever collected. Roughly half the corpus consists of unstructured, informal speech collected by volunteers, while the other half is largely unscripted speech collected in more formal settings, such as interviews and religious services. Collected in 1991-1992, the 10-million word corpus was designed to include speech from across the United Kingdom, and its annotations include speaker-specific metadata about age, sex, occupation, location, and other details of sociolinguistic relevance. The Audio BNC was originally recorded by volunteers on 1,213 90-minute cassette tapes, which were then transcribed into British English orthography by professional audio typists. These transcriptions were linguistically annotated and published as part (10%) of the British National Corpus ([11], [12]). In 2009-10, the British Library Sound Archive digitized most of the original audio recordings to stereo PCM audio (.wav files) at 96 kHz with 24-bit resolution. We downsampled these to 16-bit, 16 kHz monophonic files, and automatically aligned the orthographic transcriptions with the audio files, using the HTK speech recognition toolkit ([13]), with an HMM topology to match the Penn Phonetics Laboratory Forced Aligner (P2FA: [14]), with a combination

of P2FA American English plus our own UK English acoustic models. In accordance with the recording agreements and publication principles of the BNC transcriptions, personal names and some other speaker-specific information in the recordings were silenced to respect speaker anonymity. We have recently published the anonymized audio and time-aligned transcriptions on-line ([15]).

3. Methods

3.1. Selection of Data

From the Praat text grids generated by forced alignment, we compiled an index of word pair locations (filename, and word pair start and end times) for the entire Audio BNC. Once approximately 32,000 word pairs of interest had been located within the Audio BNC index (as described below), each token was listened to in order to exclude from further analysis all word pairs that had been grossly misaligned. Across all word pairs, the alignment of transcription and audio matches in 67% of cases: in one-third of tokens, the complete word pair was not audible in the corresponding audio clip. From these verified word pairs, the analysis was further restricted to tokens for which metadata about the speaker, specifically sex, was available. 14,000 word pairs selected for analysis were extracted from the original audio files and re-aligned automatically with a modified dictionary (see also 4.2) to improve phone boundary locations by allowing for shorter phoneme duration and alternative phoneme labels.

To identify environments for potential non-canonical assimilation, we searched the word pair index for word pairs in which the first word ends in a nasal consonant, and the second begins in an oral consonant. We restricted the study to nasals because we could measure formant frequencies during their closure portion, which is not possible with oral stops. We searched for bilabial nasal /m/ before a velar stop (e.g. *I'm gonna*, 325 tokens; *I'm going*, 424 tokens; *I'm getting*, 58 tokens), and before alveolars (e.g. *I'm trying*), as well as in control contexts in which assimilation is not expected (before another bilabial, e.g. *come back*, 374 tokens; *I'm putting*, 10 tokens). Likewise, we searched for velar nasal /ŋ/ before alveolar stops (e.g. *trying to*, *long time*) or bilabials (e.g. *coming back*, 125 tokens; *going back*, 111 tokens), and in a velar-velar control context (e.g. *dressing gown*, 14 tokens; *young girl*, 13 tokens). We also collected pairs in which the first word ends in an alveolar /n/, before labial or velar stops; these nasals are expected to assimilate.

3.2. Formant Frequency Analysis

We analyzed formant frequencies in the word-final nasal of interest and in the vowel immediately preceding it in order to infer the place of articulation of the nasal. Using Praat acoustic analysis software ([16]), we automatically measured F1, F2, F3 frequencies in the aligned word pairs. To measure formant frequencies, Praat computes a Burg spectrum with a time step of 0.0125s, an effective analysis window of 0.025 seconds, and a pre-emphasis of 50 Hz (these are Praat's standard settings). Males' and females' data were measured and analyzed separately, with the following settings: for male speakers, five formants were measured with a maximum range of 4500 Hz; for female speakers, four formants were measured with a maximum range of 5500 Hz. In order to normalize over vowels and nasal consonants of different durations, the formant frequencies were measured at 10% fractions of each segment in a word pair (0%–90%). However, as the automatically-placed segment boundaries were found to be quite accurate, and as the formant frequencies are quite stable during the vowels and nasals examined in this paper, we averaged across deciles to obtain a single mean value of each parameter for each segment.

4. Results

4.1. Accuracy of Word and Segment Boundaries Found by Forced Alignment

The recordings in this corpus are very challenging for forced alignment and formant frequency tracking. Due to the informal recording methods (Sony Walkman cassette recorders with built-in condenser microphones, used by volunteer members of the public in a wide variety of recording environments), the signal-to-noise ratio in many of the recordings is so poor that in visual examination of their spectrograms, it can be extremely difficult even for an expert to discern formants or cues to segment boundaries. Therefore, we evaluated the accuracy of word and segment boundaries assigned by the forced aligner against two reference sets of hand-corrected boundaries. For the word boundary evaluation, the absolute differences between automatic and manually-corrected times at three data points (the start and end of word 1, and end of word 2) were calculated in 549 of the highest-frequency word-pairs. 60% of the automatically-assigned boundaries were within 50 ms of the corresponding manual boundaries, and 80% within 100 ms; the RMS difference was 70 ms. For the segment boundary evaluation, the start and end time of the word-final nasals in 374 tokens of *come back*, 126 tokens of *coming back* and 99 tokens of *coming down* were examined. 50% of the automatically-assigned boundaries were within 50 ms of the corresponding manual boundaries, 65% within 70 ms and 80% within 100 ms; the RMS difference was 80 ms. Crucially, these differences had no material effect on the statistical analysis presented below, giving us confidence in the validity of using automatically aligned data. Consequently, the measurements and statistics reported below are based on the automatic alignments.

4.2. A Selection of Planned Comparisons

The Audio BNC's most frequent word pairs (e.g. *on the*, with 4632 tokens available for analysis) are also the most likely to be spoken more quickly, reduced in casual speech, and are hardest for the aligner to segment accurately. The formant frequency measures of less-frequent word pairs (e.g. *young girl*, 13 tokens) are highly variable across tokens, making statistical comparisons difficult. The clearest data come from word pairs that are frequent enough for statistically useful results, but not so frequent as to be prone to extreme phonetic shortening. We focus here on 4424 final nasals in the first word of selected pairs, in some cases pooling across multiple pairs (e.g. *from the* and *from there*). We compared the formant frequencies of labial, alveolar and velar variants as selected by the automatic aligner, which was free to assign the most-likely labels to the audio. The aligner's choice was restricted by word-pair context; for example, the nasal of *some* could be only [m] in *some people*, but could be [m] or [n] in *some time*. Word-final /ŋ/ in gerunds (see 4.4) could be [m], [n] or [ŋ].

From prior work (e.g. [17]) on acoustic cues to place of articulation, we expected that F1 would not be very different between [m], [n] and [ŋ]. We expected that [m] would have

the lowest F2 and F3, especially at the transition from the preceding vowel; that [ŋ] would have a higher F2, possibly rising in the direction of a falling F3, especially after front vowels (the "velar/palatal pinch"); and that F2 and F3 for [n] would be highest of all. We expected these differences to persist even in values averaged across decile measurements. We also expected that the aligner's usage of labels M, N and NG will reflect such acoustic differences. In fact, significant differences were only found sporadically for F3, so in the following sections we report only F2 values.

4.3. Variation in Bilabial Nasals

In this comparison we examined whether final /m/ in *come*, *from*, *I'm*, and *some* varies according to (and in the direction of) a following alveolar or dental consonant (/d, t, ð/) or velar stop (/g, k/). Table 1 presents F2 values of word-final nasals, according to the phone label assigned by the automatic aligner (M, N, or NG). Results were pooled across pairs for each word: for example, *I'm* includes the control *I'm putting*, where only [m] is expected, and the labial-velar and labial-alveolar test pairs *I'm going* and *I'm talking*.

Table 1. *Means, standard deviations, and statistical comparisons of nasal F2 frequencies, lexical /m/. M, N, NG in column 1 are aligner labels.*

Word 1	sex (*N*)	F2 in Hz (SD)	p-value
come M	M (235)	1228 (186)	0.002
come N	M (42)	1308 (163)	
come M	F (220)	1579 (274)	0.006
come N	F (67)	1667 (237)	
from M	M (422)	1157 (192)	p ≈ 0
from N	M (388)	1321 (262)	
from M	F (136)	1526 (299)	p ≈ 0
from N	F (256)	1705 (273)	
I'm M	M (282)	1277 (219)	M-N: 0.04
I'm N	M (74)	1322 (197)	N-NG: 0.008
I'm NG	M (172)	1396 (265)	M-NG: p ≈ 0
I'm M	F (321)	1585 (281)	M-N: 0.01
I'm N	F (78)	1663 (263)	N-NG: 0.003
I'm NG	F (164)	1766 (287)	M-NG: p ≈ 0
some M	M (216)	1236 (199)	M-N: 0.003
some N	M (34)	1354 (224)	N-NG: n.s.
some NG	M (28)	1325 (211)	M-NG: 0.02
some M	F (140)	1565 (311)	M-N: n.s.
some N	F (25)	1638 (304)	N-NG: 0.08
some NG	F (28)	1755 (294)	M-NG: 0.002

The means of nasal formant frequencies were compared using one-tailed t-tests within each word, across the selected nasal variants, for each sex. As expected, when these lexically labial nasals are labelled by the aligner with M, their F2 values are significantly lower than when they are labelled as N. This holds across all comparisons within *come*, *I'm*, and *from*, and across the two categories of *some* for male speakers only. When the NG label is selected for *I'm*, F2 is significantly higher than with N and M labels. In *some*, the F2 of NG variants is significantly higher than the M variants.

4.4. Variation in Velar Nasals

In this comparison we examined whether F2 of final /ŋ/ in *ring*, *long*, *young*, and *coming* varies according to the following consonant. Although gerunds such as *coming* are relatively frequent in the Audio BNC, they may also have regular [n] variants (i.e. *comin'*), as reported in the sociolinguistics literature (e.g. [18]). Because bilabial tokens might thus be regarded as assimilated forms of these alveolar variants, evidence of assimilation in these verb forms is somewhat problematic, and weaker than evidence from monomorphemes which we focus on here, such as *long*, *ring*, and *young*. Table 2 presents F2 values of word-final lexical velar nasals, according to the phone label assigned by the automatic aligner.

Table 2. *Means, standard deviations, and statistical comparisons of nasal F2 frequencies, lexical /ŋ/. M, N, NG in column 1 are aligner labels.*

Word 1	sex (*N*)	F2 in Hz (SD)	p-value
ring N	M (9)	1346 (334)	0.054
ring NG	M (31)	1558 (264)	
ring N	F (6)	1710 (174)	0.043
ring NG	F (32)	1873 (266)	
long M	M (15)	1172 (252)	M-N: n.s.
long N	M (70)	1166 (271)	N-NG: 0.015
long NG	M (133)	1084 (223)	M-NG: n.s.
long M	F (10)	1332 (170)	M-N: 0.008
long N	F (80)	1504 (322)	N-NG: n.s.
long NG	F (87)	1466 (372)	M-NG: 0.03
young M	M (25)	1180 (189)	M-N: n.s.
young N	M (9)	1227 (174)	N-NG: n.s.
young NG	M (41)	1177 (301)	M-NG: n.s.
young M	F (15)	1425 (355)	M-N: 0.01
young N	F (8)	1761 (266)	N-NG: n.s.
young NG	F (18)	1579 (334)	M-NG: n.s.
coming M	M (23)	1276 (219)	M-N: 0.004
coming N	M (31)	1446 (231)	N-NG : n.s.
coming NG	M (58)	1403 (290)	M-NG: 0.018
coming M	F (24)	1621 (239)	M-N: p ≈ 0
coming N	F (19)	1916 (217)	N-NG: 0.039
coming NG	F (69)	1813 (230)	M-NG: p ≈ 0

In *ring*, F2 was significantly lower when the label N was selected (i.e. before alveolar consonants). For males, the significance was marginal. In *long*, significant differences varied by sex: female speakers' F2 was significantly higher in tokens labelled N and NG than in those labelled M, while for male speakers F2 was unexpectedly highest in M-labelled tokens. In *young*, the F2 of N tokens was significantly higher than that of NG variants for female speakers. Velar nasals are much rarer in the Audio BNC than labials, which may contribute to the comparatively low statistical significance of these differences.

4.5. Evidence of Assimilation

Throughout Tables 1 and 2, we find acoustic differences that are expected in assimilation: for each word-final nasal, F2 is generally lowest when the M label is selected and highest when the label is NG. These findings are robust, and they hold even when data from multiple word pairs are pooled. For each word, the aligner assigns to a large number of tokens a label that does not match the lexical nasal (e.g., the >100 tokens of *come* labelled N in Table 1). Since we expected assimilation only to occur in a small proportion of cases, this prompted us to consider whether these rates reflect the actual incidence of assimilation. As a check, one of the authors (Coleman), an

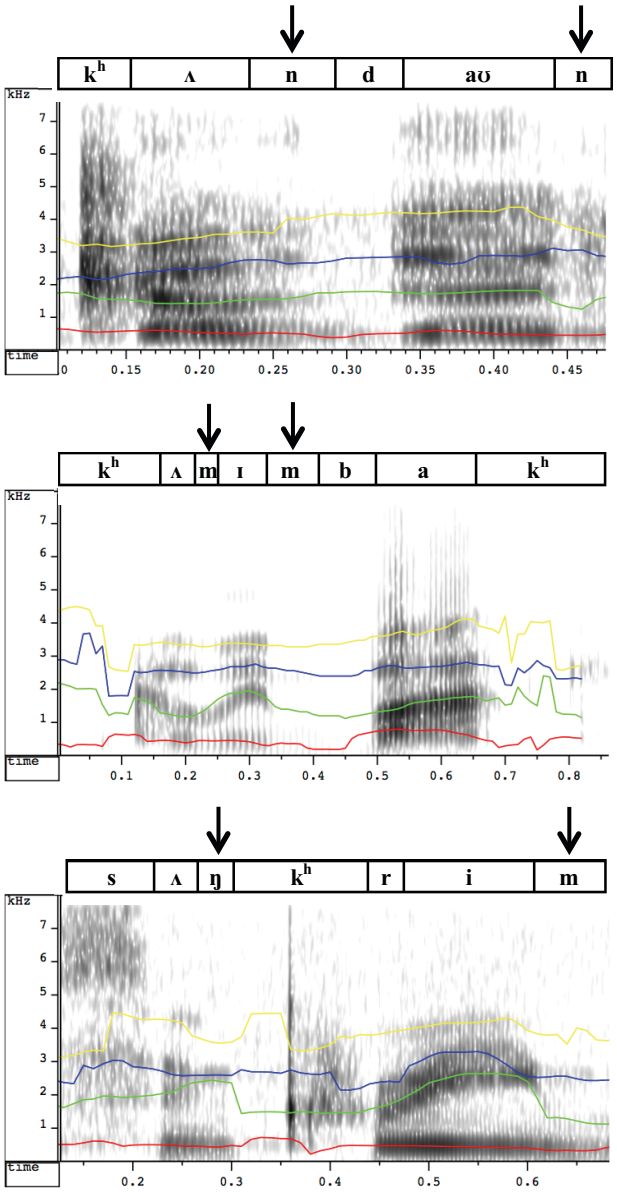

Figure 1: *Wide-band spectrograms, with overlaid formant tracks, of (top): "come down", with final nasal in "come" pronounced [n] (cf. [n] in "down"); (middle) "coming back", final nasal in "coming" pronounced [m]; (bottom) "some cream", final nasal in "some" pronounced [ŋ] (cf. [m] in "cream"). Arrows indicate nasal stops. The transcriptions were manually placed and assigned. Red line: F1 frequency, green line: F2 frequency, blue line: F3 frequency, yellow line: F4 frequency.*

experienced phonetician, listened to 668 tokens classified as assimilated by the aligner. Clear cases of assimilated non-alveolar nasals were identified; Figure 1 presents three examples of these audibly-assimilated non-alveolar nasals at word boundaries. In Figure 1 (top), the F2 of /m/ in *come down* does not drop, as in the [m] of *coming* (middle) and *cream* (bottom), but remains high, similar to the [n] in *down*. The /ŋ/ in *coming back* (middle) has a falling F2 similar to that of the medial /m/ of *coming* and the final [m] of *cream* (bottom). In *some* (bottom), the F2 of /m/ clearly rises towards F3, constituting a "velar pinch" before the /k/ of *cream*.

However, such clearly-audible assimilations are rare in the data set: in the listening test, where labials were labelled as "N", 3 of 109 tokens of *come down* and 2 out of 24 tokens of *some* were heard as alveolar; before a dental, 8 of 215 tokens of *from the* sounded unambiguously coronal e.g. [fɹən̪n̪ə] (*cf.* [19]), while 2 of 53 tokens of *from there* were audibly assimilated. Six out of 56 tokens of *some* labelled "NG" were clearly velar, while among velars labelled as "M", only 1 out of 5 tokens of *wrong* sounded labial. In summary, 3.3% of tokens in the listening test sounded definitely assimilated. These discrepancies between the aligner's labels and auditory judgments suggest that while this aligner's acoustic models perform very well at temporally identifying segments in the speech signal (4.1), it is perhaps not adapted for variant selection at the level of precision required to detect rare assimilations. Equally, however, the subjective listening test has limitations such as listener biases and difficulties with unclear or ambiguous stimuli. In future work, it would be worthwhile to employ a recognizer specifically trained for this purpose (e.g. adaptive discriminative training). Nevertheless, while audible distinctions in segment quality are not as common as the aligner's labels imply, the statistically significant differences in Tables 1 and 2 demonstrate that the variants identified by the aligner do correlate robustly with acoustic distinctions, which are in turn likely to correspond to variation in place of articulation. Besides total assimilation, other possible sources for these differences include coarticulation or gestural overlap between a word-final nasal and the consonant that follows it.

5. Conclusion

For word-final nasals, "coronal underspecification" theory ([7]) offers an unambiguous prediction: labials and velars are not expected to assimilate their place of articulation to that of following consonants. We tested this prediction against over 4,000 relevant word-pairs from a corpus of mostly unscripted, spontaneous English speech. We found strong evidence that word-final velar and bilabial nasals sometimes assimilate to following consonants (e.g. Figure 1 and other such clear cases). The strongest evidence came from the more numerous word pairs, such as *come down* (190 tokens), while in word pairs with fewer tokens such as *long time* (84 tokens), assimilation was not unambiguously observed. Though we cannot yet reliably detect assimilation automatically, the clear cases we do have of assimilation by bilabial and velar nasals mean that phonological theory needs to be revised. We think that a probabilistic approach to phonology (e.g. [20]) could model such assimilation patterns more adequately. In such a model, alveolars, velars and bilabials could *all* assimilate, but with different ranges of variation and with different incidences, determined for example by place of articulation and relative frequency of nasal word-pairs.

6. Acknowledgement

This work was supported by the UK Economic and Social Research Council, grant reference RES-062-23-2566.

7. References

[1] Kriedler, C. W. The Pronunciation of English. Oxford: Blackwell, 1989.

[2] Harris, John. English Sound Structure. Oxford: Blackwell, 1994.

[3] Roca, I., and Johnson, W. A Course in Phonology. Oxford: Blackwell, 1999.

[4] McMahon, A. An Introduction to English Phonology. Edinburgh: Edinburgh University Press, 2002.

[5] Shockey, L. Sound Patterns of Spoken English. Malden, MA: Wiley Blackwell, 2003.

[6] Cruttenden, A. Gimson's Pronunciation of English. London: Hodder Education, 2008.

[7] Avery, P., Rice, K. "Segment structure and coronal underspecification", Phonology 6:179-200, 1989.

[8] Dilley, L. C., Pitt, M. A. "A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition", J. Acoust. Soc. Am. 122(4):2340-2354, 2007.

[9] Ogden, R. "A declarative account of strong and weak auxiliaries in English", Phonology 16:55-92, 1999.

[10] Lodge, K. A Critical Introduction to Phonetics. London: Continuum, 2009.

[11] Crowdy, S. "The BNC spoken corpus", *in* Leech, G., Myers, G., Thomas, J., eds. Spoken English on Computer: Transcription: mark-up and application, London: Longman, 224-234, 1995.

[12] BNC Consortium. "BNC XML Edition." DVDs from http://www.natcorp.ox.ac.uk/, 2007.

[13] Young, S., Evermann, G., Gales, M., and 9 others. "The HTK Book (for HTK Version 3.4)." http://www.ee.uwa.edu.au/~roberto/research/speech/local/htk/htkbook.pdf, 2009.

[14] Yuan, J., Liberman, M. "Speaker identification on the SCOTUS corpus." Proceedings of Acoustics '08, 2008. Software at: http://www.ling.upenn.edu/phonetics/p2fa/

[15] Coleman, J., Baghai-Ravary, L., Pybus, J., Grau, S. "Audio BNC: the audio edition of the Spoken British National Corpus." Phonetics Laboratory, University of Oxford, http://www.phon.ox.ac.uk/AudioBNC, 2012.

[16] Boersma, P., Weenink, D. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/, 2012.

[17] Olive, J. P., Greenwood, A., Coleman, J. Acoustics of American English Speech. New York: Springer-Verlag, 1993.

[18] Trudgill, P. The Social Differentiation of English in Norwich. Cambridge: Cambridge University Press, 1974.

[19] Manuel, S. Y. "Speakers nasalize /ð/ after /n/, but listeners still hear /ð/", Journal of Phonetics 23:453-476, 1995.

[20] Jun, J. "Place assimilation", *in* Hayes, B., Kirchner, R., Steriade, D., eds. Phonetically Based Phonology. Cambridge: Cambridge University Press, 58-86, 2004.