On the presence of final lowering in British and American English

Amalia Arvaniti

Department of Linguistics,
University of California, San Diego,
9500 Gilman Drive #0108,
La Jolla, CA, 92093–0108, USA
amalia@ling.ucsd.edu

Abstract

Lists of bean names with two to five items were elicited from speakers of Mainstream American English (MAE) and Standard British English (SBE), and three methods for detecting final lowering in these data were used: comparisons of the scaling of final and penultimate peaks which have the same order in the lists (e.g., third peak), modelling of the data as exponential decay followed by comparisons of attested and predicted final peak scaling, and comparison of interaccent drops in F_0. The first two measures showed that final lowering was present in both MAE and SBE, while the comparison of F_0 drops across peaks showed only very weak evidence for final lowering. Further, the results showed that final peak scaling was not affected either by durational differences in the interval between penultimate and final peaks, or by the distance of the final peak from the end of the utterance. Together, these results suggest that final lowering is independent of declination and targets the final accent of utterances, indicating that final lowering is grammaticalized in the linguistic varieties examined here. Finally, the differences between the methods used to detect final lowering show that its effect on peak scaling is very small and thus caution is needed when choosing a method for its detection and when interpreting the results.

Keywords: declination, final lowering, tonal scaling, downstep

Languages: American English, British English, Japanese, Danish, Dutch, Yoruba, Kipare, Spanish, German

1. Introduction

Final lowering, the lower than expected scaling of the final peak in a series of downstepping H tones, was first reported by Liberman and Pierrehumbert (1984). Liberman and Pierrehumbert (1984) recorded lists of two to five berry names (e.g., *raspberries, bayberries and mulberries*) at three different pitch ranges and used their materials to model the scaling of peaks. They showed that the scaling of each peak in these contours could be modelled was a fraction of the scaling of the preceding peak, with values reaching a speaker and range specific reference level that is higher than zero. The models created for each of Liberman and Pierrehumbert's three speakers showed very small differences between predicted and attested peak values, but they also consistently predicted much higher values for final peaks than those attested, with differences ranging from 4 Hz to 27 Hz depending on speaker and pitch range.

Liberman and Pierrehumbert (1984) dubbed this phenomenon "final lowering" but remained agnostic as to its origins, suggesting that it could either have a "physiological explication" or be "under phonological control" (Liberman and Pierrehumbert 1984: 219). They further suggested that if final lowering is a phonological phenomenon it would be more likely to target the last accent in an utterance, independently of the position of this accent relative to the end of the utterance. If, on the other hand, final lowering is a physiological phenomenon, it would be more likely to target a stretch at the end of the utterance. Possible physiological explanations for final lowering offered by Liberman and Pierrehumbert include a drop in subglottal pressure (which they found unlikely), and a relaxation of the laryngeal muscles towards the end of the utterance. More recently Herman, Beckman and Honda (1996) and Herman (2000) have provided data compatible with both physiological explanations. Specifically, Herman et al. (1996) and Herman (2000) suggest that final lowering is one of the several manifestations of overall "vocal effort" decrease towards the end of an utterance. Less vocal effort results in a drop in subglottal pressure and in gesture stiffness, including a relaxation of the muscles of the larynx; these changes can lead to segmental lengthening, a drop in average (Root Mean Square) amplitude, voice quality changes and final lowering. Further, Herman (2000) shows that each of the above phenomena can be used as an indicator of discourse-finality; e.g., in her corpus, identical sentences with the same intonational and

prosodic structure showed lower final F_0 peaks when they were in discourse-final position than discourse-medially. Under this interpretation of the function of final lowering, the data of Liberman and Pierrehumbert (1984) could be showing that their speakers treated each of their utterances as discourse final.

Evidence for final lowering has been presented for a variety of other languages with very different prosodic systems, such as Japanese (Poser 1984; Pierrehumbert and Beckman 1988), Danish (Thorsen 1985), Dutch (Gussenhoven and Rietveld 1988), Yoruba (Connell and Ladd 1990; Laniran 1992), Kipare (Herman 1996) and Spanish (Prieto, Shih and Nibert 1996). Documenting final lowering in such a diverse group of languages supports the view that final lowering is a physiological phenomenon. However, a closer look also reveals that its application is not exceptionless. As mentioned, in American English final lowering is used only with discourse-final utterances (Herman 2000); similarly, Herman (1996) shows that in Kipare final lowering is used in declaratives and yes-no questions but not in sentences meant to show incredulity. This meaningful variation in the application of final lowering indicates that—in some languages at least—final lowering is not a "phonologically irrelevant consequence of the way utterances are ended" (Liberman and Pierrehumbert 1984: 219); rather, it appears that although there are good physiological reasons why final lowering takes place, the phenomenon is now grammaticalized in at least some languages; consequently, it is likely to show variation in function and applicability to particular tunes, as well as cross-linguistic and cross-dialectal variation.

Despite the widespread evidence for final lowering, its very existence in English has been called into question by the results of Grabe (1998). Grabe measured the F_0 difference between successive peaks in German and British English materials of a structure similar to that of Liberman and Pierrehumbert (1984). Her English results showed that when the distance (in syllables) between successive peaks was kept constant, the step in F_0 between the first and subsequent peaks was larger than any of the following steps, between which no statistically significant differences were found. This result, which was interpreted as lack of final lowering, prompted Grabe to suggest that the final lowering found in Liberman and Pierrehumbert's study may be an artefact of their experimental design. Specifically, in the materials of that

study the last two items in the lists were always separated by *and* (e.g., *blueberries, bayberries, mulberries <u>and</u> raspberries*), the presence of which resulted in the interval between the last two accents (henceforth *interaccent interval*) being longer by one syllable than all preceding intervals. This difference led Grabe to hypothesize that the final lowering observed in Liberman and Pierrehumbert (1984) was due to the greater declination between the last two accents in their materials, and not an independent phenomenon. Grabe's own results from a follow-up experiment support this interpretation by showing that accents separated by more segmental material from a preceding accent show lower scaling than accents which are closer to a preceding accent (e.g., the accent of *moonbeam* is scaled lower in *moonlighting, moonbeam* than in *moonlit, moonbeam*).

However, other explanations for the discrepancy between the results of Liberman and Pierrehumbert (1984) and Grabe (1998) are also possible. To begin with, the fact that final lowering is present in the former study but not in the latter is an indication that it may be an optional component of tunes. It is thus possible that the speakers in Grabe (1998) used a tune (or tunes) that did not involve final lowering. There are two possible reasons for this. First, Grabe's subjects spoke Standard British English, while Liberman and Pierrehumbert's data were elicited from American speakers; it is possible that these varieties employ different tunes for lists and that only the tune used by American speakers involves final lowering (for dialectal differences in intonation see e.g., Grabe, Post, Nolan and Farrar 2000; Atterer and Ladd 2004; Arvaniti and Garding in press; Prieto, D'Imperio and Gili Fivela in press; Dalton and Ní Chasaide, this volume). On the other hand, it is also possible that the differences between the studies are due to the materials used rather that then tunes. Specifically, Grabe's materials did not include *and* between the last two items; if *and* signals to the speakers that they are reaching the end of a discourse unit in which final lowering is required (and assuming that final lowering has the same discourse function in British English as in American English), then lists in which the last two items are not connected with *and* may not serve the same discourse function as lists in which they are.

The experiment presented here was designed to test some of these hypotheses regarding the nature of final lowering in English and the extent to which it is used in Mainstream American English (henceforth *MAE*) and (Southern) Standard British English (henceforth *SBE*).

2. Method

2.1. Materials

The materials were similar, though not identical, to those of Liberman and Pierrehumbert (1984). First, bean names (rather than berry names) were used as they made it possible to vary the number of syllables in the first member of the compounds and by doing so, manipulate the syllable count between accents. To this purpose, seven bean names were selected to create lists with two to five items; these test items are shown in Table 1 below.

Table 1: Test-words used in the experiment

two syllable	three syllables	four syllables
green beans	*lima beans*	*haricot beans*
long beans	*yellow beans*	
	fava beans	
	navy beans	

Second, in Liberman and Pierrehumbert (1984) each test-word appeared in all possible positions in a list as a way of controlling, among other things, for differences in the intrinsic F_0 of the accented vowels (on intrinsic F_0 see, e.g., Whalen and Levitt 1995, and Connell 2002). Here the order of the test-words was used to control the size of interaccent intervals, and because of this requirement each test-word appeared in a single position in all lists.[1] Specifically, the number of unstressed syllables between accents was two in all positions, except between the penultimate and final accent. As shown in (1) and (2) below, in penultimate position either *green beans* or *navy beans* was used. When *green beans* is in penultimate position, as in (1), *and* does not add an extra syllable between the last two accents, and thus all interaccent intervals in the list are of equal length, as shown in the underlined part

of (1); in this way the "flaw" in Liberman and Pierrehumbert's experiment is corrected. When *navy beans* is in penultimate position, as in (2), the last interaccent interval is three syllables long, as the underlined part of (2) shows; in this way the "flaw" in Liberman and Pierrehumbert's design is replicated. These materials allow for testing the declination hypothesis: if final lowering is due to declination, then only lists with *navy beans* in penultimate position should show final lowering; if, on the other hand, final lowering is independent of declination, then both lists with *green beans* and lists with *navy beans* in penultimate position should exhibit final lowering. As both declination and final lowering could be exerting an influence on the final peak, a third possibility is that both lists with *green beans* and lists with *navy beans* in penultimate position would exhibit final lowering, but the effect would be stronger in the latter, since in these lists it is compounded by additional declination.

(1) *Lima beans, yellow beans, green <u>beans and</u> long beans.*

(2) *Lima beans, yellow beans, na<u>vy beans and</u> long beans.*

In addition, two different test-words, *long beans* and *haricot beans*, were used as the final item in the lists, as shown in examples (3) and (4). If final lowering targets the last accent, then sentences ending in either word should have final peaks of similar scaling; on the other hand, if final lowering targets the last portion of the utterance, then sentences ending in *haricot beans* would be more likely to have higher final peaks. According to Liberman and Pierrehumbert (1984) the former outcome would suggest that final lowering is phonological, while the latter would suggest that final lowering is a physiological phenomenon.

(3) *Lima beans, green beans and long <u>beans</u>.*

(4) *Lima beans, green beans and h<u>aricot beans</u>.*

The combination of penultimate and final word gave four possibilities for the final part of each list length; these are schematically presented in Table 2.

Table 2: Combinations of last interaccent interval and position of last accent from utterance end.

Distance of last accent from end	Last interaccent interval length	
	2 syllables	3 syllables
1 syllable	*(…) green beans and long beans.*	*(…) navy beans and long beans.*
3 syllables	*(…) green beans and haricot beans.*	*(…) navy beans and haricot beans.*

Originally the materials also included lists in which *and* was not used between the last two items. The aim of these materials was to test for the possibility that the difference between the results of Grabe (1998) and those of Liberman and Pierrehumbert (1984) was due to the presence or absence of *and*, and what it signals to the speakers regarding the discourse function of the utterances. However, initial recordings with MAE speakers showed that they were very unhappy with these materials; some found them so unnatural they refused to read them altogether. For this reason, lists that did not include *and* between the last two items were dropped from the experiment.

2.2. Speakers

The materials were elicited from six MAE and six SBE speakers. The speakers in both groups were in their twenties or thirties and had no known speech or hearing problems. They were all naïve as to the purposes of the experiment and had no training in intonation.

The male MAE speakers, AH, BS and JB, were from Pennsylvania, Boston, MA and Boise, ID respectively; the two female speakers whose data are reported here, AS and LN, were from Wisconsin and upstate New York respectively; the data of the third female speaker were discarded because they showed a great deal of variation in terms of speaking rate, melodies and phrasing used. Although the MAE speakers were from a geographically disperse area that includes different well-known dialectal regions (e.g., Labov, Ash and Boberg in press, chap. 11) they were selected because they were considered by several members of their immediate community (the graduate student body at the UCSD Linguistics Department) to have no strong regional accent.

The female SBE speakers RL and SF were from Hampshire and London respectively; the third female speaker, AB, and male speaker PM were from South Yorkshire; DE was from the

West Midlands and JR from Birkenhead in Merseyside. The SBE speakers were also from different regions within the UK, some of which, like Merseyside, are associated with strong regional accents (e.g., Trudgill 1999); however, they all reported speaking SBE and were assessed to do so by the experimenter. Five of the SBE speakers were at the time of the recording visiting San Diego for short periods of time, ranging from a few weeks to three months; only one speaker, JR, was a member of the UCSD faculty and had lived in the United States for several years.

2.3. Procedures

Recordings took place in the sound-treated recording booth of the UCSD Phonetics Laboratory. The MAE data were recorded on DAT tape, and were re-digitized at 16 KHz with 16-bit resolution for analysis. The SBE data were recorded directly to disk using an AD converter at a sampling rate of 44.1 KHz with 16-bit resolution.

The speakers were told to read the lists at their natural pace as if they were answers to questions such as *Which kinds of beans do you like?* or *What did you buy at the health food store?* They were not asked to use different pitch ranges, and no instructions were given to them as to what intonation or phrasing pattern to use. The materials were typed on flashcards, one list per card. The speakers read the materials ten to twelve times in random order (the number of repetitions was lower for more fluent speakers). On average eight repetitions for MAE and ten repetitions for SBE speakers were used for measurement. Tokens were discarded if they were disfluent, were produced with the wrong intonation contour altogether (e.g., with narrow focus on one of the items), or with such extensive creak that the last peak could not be measured. In general the SBE and MAE speakers used essentially the same tune, which can be autosegmentally represented as a sequence of H*+L pitch accents followed by L– L% phrasal tones.[2] They typically phrased each list as one intonational phrase (with one exception, discussed in more detail in section 3.4.).

2.4. Measurements and statistical analysis

Measurements were taken by simultaneous inspection of the waveform, wide-band spectrogram and F_0 track of each sentence using PRAAT, and included the following:

- The F_0 of all the peaks in each list (P1–P5), with peak being defined as the highest F_0 value within each word in the list (but see immediately below); in cases where there were more than one point with the same F_0, the first point was selected. These criteria applied to all peaks in both the MAE and SBE data, except for the first peak, which was aligned differently in the two varieties. As can be seen in Fig. (1), in MAE, P1, which typically occurred early within the vowel of *beans* in *lima beans*, was measured in the same way as the other peaks; in SBE, however, in which the first peak aligned towards the end of *beans*, P1 was taken to be the highest F_0 point within *lima*. This difference in measurement, which resulted in P1 being in similar location in the SBE and MAE contours, was considered necessary in order to keep the MAE and SBE data comparable and the interaccent intervals similar in duration. This difference is unlikely to affect the results presented here, which deal primarily with the scaling of penultimate and final peaks. [3]

- LL (*Last Low*), the lowest F_0 value that could reliably be measured excluding regions of creak.

- The temporal distance between successive peaks.

The data were analyzed statistically by means of repeated measures and between subjects analyses of variance (ANOVA), depending on the comparison in question. Planned comparisons were used to test for differences between the levels of a factor and to explore expected interactions (no analysis yielded unexpected interactions). Mean speaker values rather than raw data were used in these ANOVAs in order to avoid violating the independence of observations (one of the main assumptions underlying ANOVA), and an increase in type I error, due to the large number of degrees of freedom for the error term associated with the use of raw data (these considerations apply whether raw data are tested in separate ANOVAs for each speaker or in the same ANOVA, with speaker as a between subjects factor; for a thorough discussion of these problems see Max and Onghena 1999).
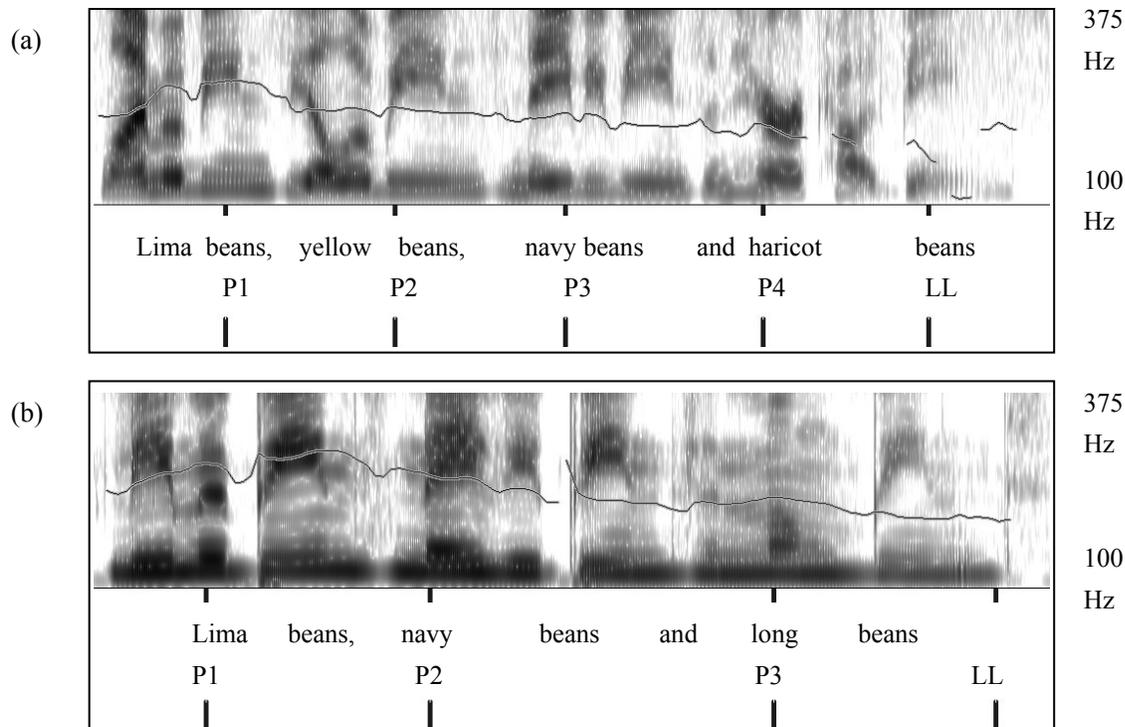
Fig. 1: Illustration of measurements; in panel (a), a four-word list from MAE speaker AS; in panel (b) a three-word list from SBE speaker RL. The difference in the quality of the first vowel visible in the spectrogram is due to the fact that the MAE speakers pronounced *lima* as [laɪma], while the SBE speakers pronounced it [lima].

Before being used for statistical analyses, the F_0 data were converted from Hz to ERB (Equivalent Rectangular Bandwidth) applying the formula of Hermes and van Gestel (1991: 97) given in (5) below.

(5) ERB = 16.7 log (1 + f / 165.4)

where f is frequency in Hz. This was done in order to reduce pitch range differences across speakers and genders whenever data were pooled. On the other hand, F_0 values in Hz were used for the modelling of individual speaker data (see section 2.5).

2.5. Methods of testing for final lowering

Three methods were used to detect final lowering. First, following Liberman and Pierrehumbert (1984), the scaling of the final peak in word lists with n accents was compared

to the scaling of the penultimate peak of word lists with *n*+1 accents: if final lowering is present, then the former peak, being final, should be scaled lower than the latter.

This method has the disadvantage that it can attribute to final lowering scaling differences which are due to other factors. In particular, if a speaker regularly produces a list as two phrases, rather than one, it is likely that she will also have declination reset at the start of the second phrase (Cooper and Sorensen 1981; O'Shaughnessy and Allen 1983; Ladd 1988; van den Berg, Gussenhoven and Rietveld 1992). If such reset takes place it can seriously alter the scaling relations between peaks (see, e.g., PM's data in Fig. 3). Because of this possibility, the data were inspected for phrasing and declination reset, and only those in which reset was absent were used with the two methods presented below.

The second method used for detecting final lowering was data modelling that allowed for the comparison of attested and predicted final peak scaling. The modelling undertaken here was based on Liberman and Pierrehumbert (1984) and Prieto et al. (1996) with some small modifications. Specifically, the scaling of peaks in four- and five-word lists was modelled as exponential decay to a non-zero asymptote, using the following equation:

(6) $P = as^X + r$

This exponential function was converted into the linear equation in (7)

(7) $\ln(P{-}r) = \ln(s)x + \ln(a)$

where *r* is a reference line (i.e., a value starting 30 Hz below the attested mean value of the final peak for a given speaker), *x* is the order of the peak in the list (1–4 or 1–5), and *s* is a constant that is smaller than 1 and represents the amount of downstep for a given list length. Linear regression with values of *r* increasing in steps of 0.1 Hz was performed on equation (7), and the line with the highest correlation coefficient was chosen as the one best fitting the data.

Once the best-fit model was selected, the presence of final lowering was determined by calculating the difference between attested and predicted values for final peaks. Positive values clearly show that final lowering is not present (since in this case attested values are

higher than predicted values). On the other hand, a decision had to be made as to how much of a negative difference should be considered sufficient for showing final lowering. One possible way to resolve this issue is to take as evidence for final lowering any differences between predicted and attested values that would be perceptible (on the assumption that hearers can make a paradigmatic comparison between what they would expect if a peak was non-final and what they hear when the peak is final). Reported difference limens for F_0 vary to a great extent depending on the study, task and type of stimuli used (for reviews see Beckman 1986; 't Hart, Collier and Cohen 1990). Here a relatively conservative approach was taken to use a difference of 3 Hz, established by Klatt (1973) as the largest limen needed to hear a difference between two stimuli with changing F_0.[4]

Finally, the third method used for the detection of final lowering was that employed by Grabe (1998). Specifically, the drops in peak scaling, that is the differences in peak scaling between successive peaks, in three-, four- and five-word lists were calculated and compared. The aim was to see if final steps in F_0 drop would be larger than preceding steps, a result that according to Grabe (1998) would indicate final lowering.

3. Results

3.1. Pitch range

In order to ensure that any comparisons of scaling were meaningful, it was necessary to establish that each speaker used a relatively stable pitch range during his/her recording session, and that these ranges were comparable across speakers and dialects. Three measurements were used to assess pitch range, the scaling of P1 (the first peak in each utterance), the scaling of LL, and the scaling difference between them. Speaker and grand means and standard deviations (in ERB) are presented in Table 3. These data suggest that the pitch ranges used were relatively stable within the data of each speaker—in that they mostly showed small standard deviations—and comparable across dialects—in that the P1, LL and P1–LL data of the SBE speakers were not different from those of the MAE speakers, according to ANOVAs with DIALECT and GENDER as between subjects factors [F(1,7)=0.9, F(1,7)=0.9, F(1,7)=1.8, for P1, LL and P1–LL respectively; all n.s.]. (Not surprisingly, the male speakers had lower P1

and LL values than the female speakers [for P1, $F(1,7)=41.23$; for LL, $F(1,7)=160.01$; $p<.01$ for both] but their ranges were not significantly different [$F(1,7)=3.04$, n.s.]; there was no interaction between DIALECT and GENDER.]

Table 3: F_0 means in ERB and standard deviations (in brackets) for P1 (first peak), LL (last L) and pitch range (P1–LL), for each speaker separately; the gender of each speaker is given in italics next to their initials.

	SBE				MAE		
Speaker	P1	LL	P1–LL	Speaker	P1	LL	P1–LL
AB *f*	5.55(0.18)	4.76(0.59)	0.79(0.63)	AH *m*	4.76(0.44)	3.06(0.12)	1.70(0.47)
DE *m*	4.53(0.15)	3.45(0.28)	1.08(0.32)	AS *f*	6.93(0.25)	4.67(0.54)	2.26(0.51)
JR *m*	3.99(0.28)	2.9(0.028)	1.09(0.28)	BS *m*	5.04(0.31)	3.13(0.35)	1.91(0.43)
PM *m*	4.69(0.24)	2.98(0.07)	1.72(0.25)	JB *m*	4.31(0.16)	3.41(0.33)	0.90(0.35)
RL *f*	7.25(0.30)	5.06(0.45)	2.19(0.56)	LN *f*	6.85(0.22)	4.53(0.80)	2.32(0.73)
SF *f*	6.86(0.16)	4.96(0.09)	1.9(0.31)				
Grand Mean	5.57(1.27)	4.06(0.99)	1.51(0.65)	Grand mean	5.55(1.13)	3.74(0.84)	1.81(0.72)

3.2. The role of interaccent interval and final list item

Once it was established that there were no serious pitch range disparities across speakers and dialects, the data were examined for effects of final list item (*haricot beans* or *long beans* in final position) and last interaccent interval length (*navy beans* or *green beans* in penultimate position). These comparisons were undertaken first because any effects of these two factors on final peak scaling could render the pooling of data in further analyses inadvisable.

In order to test for last interaccent interval effects on duration, an ANOVA was run with DIALECT as a between subjects factor, and two repeated measures factors: LAST INTERACCENT INTERVAL (two or three syllables, i.e., *green beans* or *navy beans* in penultimate position) and LIST LENGTH (two, three, four or five words). The aim of this analysis was to test the hypothesis that the addition of a syllable to the last interaccent interval (that is using *navy beans* instead of *green beans* in penultimate position) would result in increased duration for this interval.

The results showed that the duration of the last interaccent interval increased by 27 ms when it included three rather than two syllables [$F(1,9)=9.3$, $p<.01$], that is when the

penultimate word was *navy beans*. The duration of the last interaccent interval was also affected by LIST LENGTH [F(3,27)=16.3, *p*<.01]; planned comparisons showed that the last interaccent interval was shorter in two-word lists than in all longer lists [*p*<.01 for all pair-wise comparisons], while there were no differences among the longer lists. The shorter last interaccent interval in two-word lists was most probably due to the later alignment of the first peak, which shortens the duration of this interval in two-word lists. In addition, there was no effect of DIALECT and no interactions. Mean durations and standard errors of the last interaccent interval duration for different list lengths are presented in Table 4.

Table 4: Means in ms and standard errors (in brackets) for last interaccent interval duration as a function of interaccent interval length in syllables.

		2-word lists	3-word lists	4-word lists	5-word lists
interaccent	2 syllables	526 (27)	573 (31)	593 (42)	597 (35)
interval	3 syllables	559 (25)	612 (23)	622 (4)	624 (35)

To test for the effects of interaccent interval length and choice of final list item on the scaling of the last peak, ANOVAs were run with DIALECT as a between subjects factor, and two repeated measures factors: LAST INTERACCENT INTERVAL (two or three syllables, i.e., *navy beans* or *green beans* in penultimate position) and FINAL LIST ITEM (*long beans* or *haricot beans* in final position). The aim of these analyses was to see whether the scaling of the peak of *haricot beans* and *long beans* was affected either by its distance from the preceding peak (LAST INTERACCENT INTERVAL) or by its distance from the end of the utterance (FINAL LIST ITEM).

LAST INTERACCENT INTERVAL did not affect final peak scaling even though it had resulted in longer temporal distance between the last two peaks [for two-word lists, F(1,9)=0.002; for three-word lists, F(1,9)=1.9; for four-word lists, F(1,9)=0.08; for five-word lists, F(1,9)=0.9; all n.s.]. In addition, final peak scaling was not affected by FINAL LIST ITEM, except in five-word lists in which the peak of *haricot beans* was scaled higher than the peak of *long beans* by 4 Hz (0.17 ERB) on average [for five-word lists, F(1,9)=11.3, *p*<.01; for two-word lists, F(1,9)=2.9, n.s.; for three-word lists, F(1,9)=2.2, n.s.; for four-word lists, F(1,9)=4.7, n.s.]. There was no effect of DIALECT and no interactions. Table 5 summarizes these results.

Since there was a difference in the scaling of the final peaks in five-word lists, additional

testing was undertaken to determine whether the peak of *haricot beans* was further away from the end of the utterance than the peak of *long beans*. To this effect, an ANOVA was performed on the distance between P5 and LL with final list item (*haricot* or *long beans*) as a repeated measures factor. The results showed that the peak of *haricot beans* appeared earlier than the peak of *long beans* [F(1,10)=7.9, *p*<.02], occurring 340 ms before LL on average, while the peak of *long beans* occurred 246 ms before LL.

Table 5: Mean final peak scaling in ERB as a function of last interaccent interval length in syllables (top) and final list item (bottom). Means preceded by asterisks showed significantly higher scaling in the *haricot beans* subset (see text for details). Standard errors in these data ranged from 0.46–0.5.
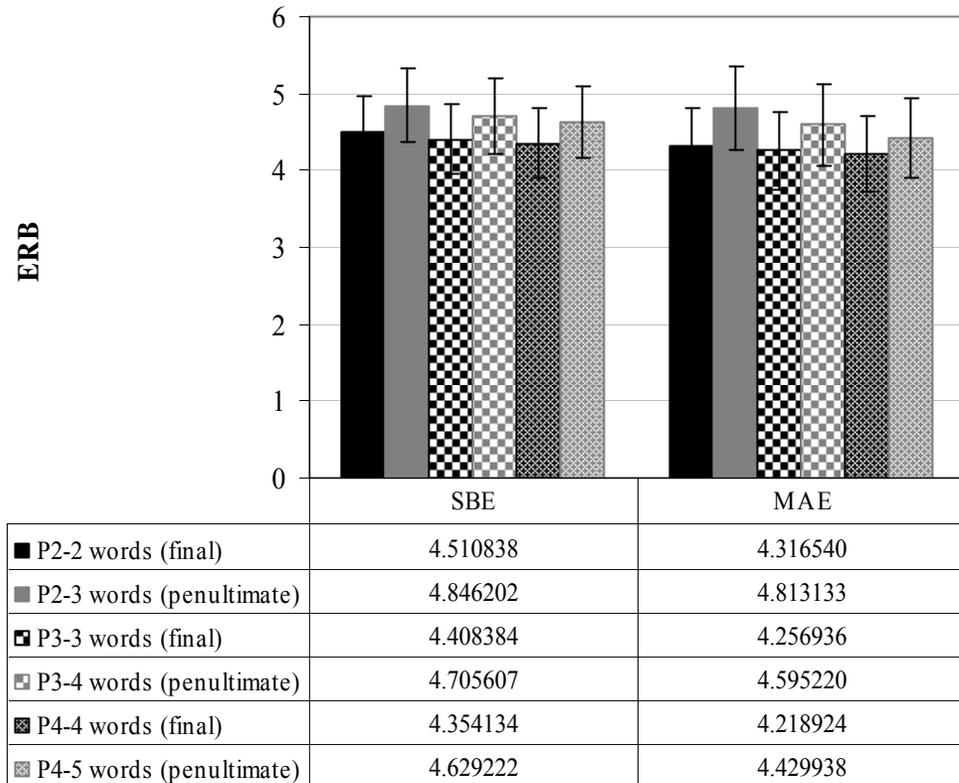
		2-word lists	3-word lists	4-word lists	5-word lists
interaccent	2 syllables	4.43	4.33	4.28	4.26
interval	3 syllables	4.43	4.34	4.29	4.27
final list	*haricot beans*	4.46	4.37	4.33	*4.31
item	*long beans*	4.39	4.3o	4.25	*4.21

Taken all together, the above results suggest that final peaks that are further away from the end of the utterance may be scaled higher that final peaks found closer to the end (the *haricot-long* difference in five-word lists), a result that according to Liberman and Pierrehumbert (1984) could indicate that final lowering is purely physiological. However, this effect was only present in the five-word lists and even in these lists both early peaks (*haricot beans*) and late peaks (*long beans*) appear well within the suggested 500 ms span of final lowering (Pierrehumbert and Hirschberg 1990), and are subject to it as will be shown below (see sections 3.3. and 3.4.). Further, the data show that even though the number of syllables in the last interaccent interval (two or three, depending on whether the penultimate item was *green beans* or *navy beans* respectively) leads to a small but consistent increase in duration, this lengthening is not accompanied by a concomitant decrease in the scaling of the last accent. Consequently, if final lowering is present, it cannot be explained away as a by-product of greater declination.

3.3. Evidence for final lowering (i): comparing final and penultimate peaks

The scaling of final and penultimate peaks was compared by means of ANOVAs with DIALECT as a between subjects factor and with ACCENT POSITION (final vs. penultimate peak)

and ACCENT ORDER (second, third, or fourth peak) as repeated measures factors; the data-points were speaker means for the relevant peaks (in ERB), e.g., P2 in two-word lists, and P2 in three-word lists. The data were pooled across different final interaccent intervals and final lexical items, since these manipulations did not apply to penultimate peaks.

	SBE	MAE
■ P2-2 words (final)	4.510838	4.316540
■ P2-3 words (penultimate)	4.846202	4.813133
◨ P3-3 words (final)	4.408384	4.256936
◩ P3-4 words (penultimate)	4.705607	4.595220
▨ P4-4 words (final)	4.354134	4.218924
▨ P4-5 words (penultimate)	4.629222	4.429938

Fig. 2: Mean scaling (in ERB) and standard errors for final peaks in two-, three- and four-word lists compared to penultimate peaks in three-, four- and five-word lists respectively. Data pooled across interaccent interval lengths and final list item, but presented separately for each dialect. Black (and white) columns represent final peaks; grey (and white) columns represent penultimate peaks.

As can be seen in Fig. 2, final peaks were scaled lower than penultimate peaks [$F(1,9)=47.2$, $p<.01$] and earlier peaks were scaled higher than later peaks [$F(2,18)=22.8$, $p<.01$; $p<.01$ for pair-wise comparisons of second vs. third peak, second vs. fourth peak, and third vs. fourth peak]. There was no effect of DIALECT and no interactions. The differences between final and penultimate peaks averaged 14.3 Hz, or 0.6 ERB (ranging from 9 Hz (0.4 ERB) for MAE fourth peaks to 22 Hz (0.9 ERB) for MAE second peaks). This method of testing for final lowering suggests that final lowering was present in the data of both SBE and

MAE speakers.

3.4. Evidence for final lowering (ii): predicting the value of final peaks

As discussed in section 2.5., in order to examine the course of peak scaling, the data were prepared for visual inspection by plotting the scaling of each peak separately for each list length and speaker. These descriptive data are presented below in Figs 3 and 4 for SBE and MAE respectively. The data from the five-word lists are highlighted so they can be more easily compared to the course of F_0 presented in Liberman and Pierrehumbert (1984) and shown in the lower right part of Fig. 4.
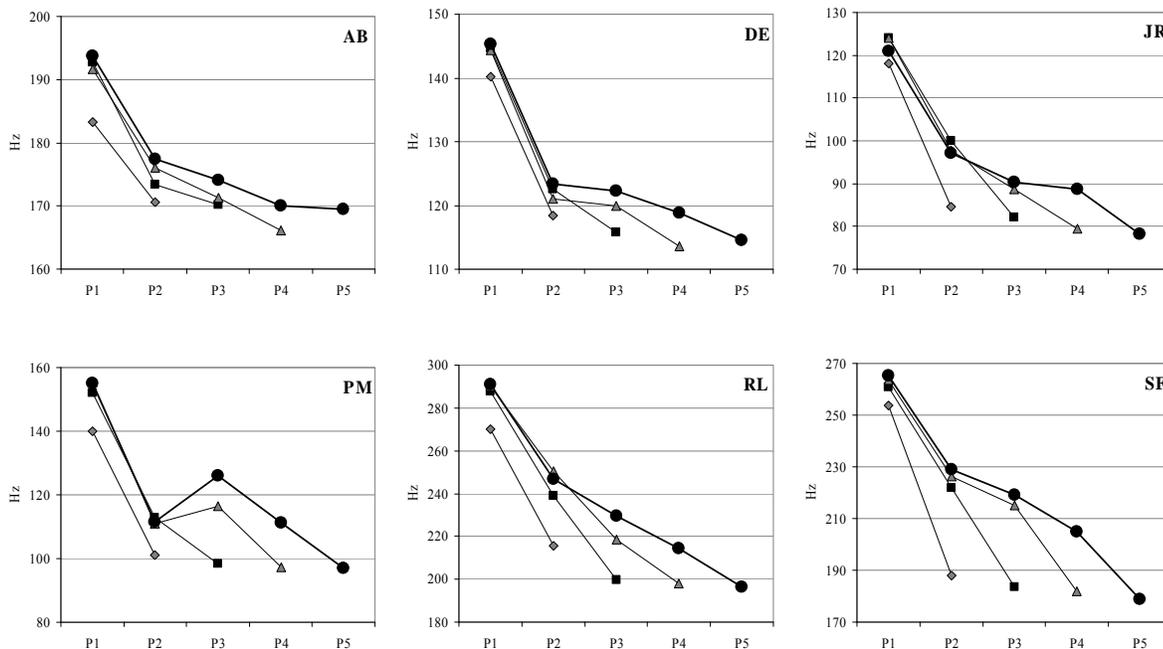

Fig. 3: Mean peak values for each of the SBE speakers, presented separately for each list length: diamonds are used for peaks in two-word lists, squares for peaks in three-word lists, triangles for peaks in four-word lists, and circles for peaks in five-word lists.

As can be seen in Fig. 3, there is evidence for pitch reset at P3 in the data of SBE speaker PM, for both four-word and five-word lists, with five-word lists, in which three accents are included in the second phrase, showing greater reset than four-word lists, in which two accents are included [F(1,39)=42.9, p<.01 according to an ANOVA on the scaling of P3 in PM's data with LIST LENGTH (four or five words) as a repeated measures factor]. Because of the phrasing

17

and declination reset, PM's data were excluded from further analysis.[5] No such problems were evident in the data of the other speakers, suggesting that they typically phrased all lists as one intonational phrase (a result that supports the impressionistic analysis mentioned in section 2.3.).
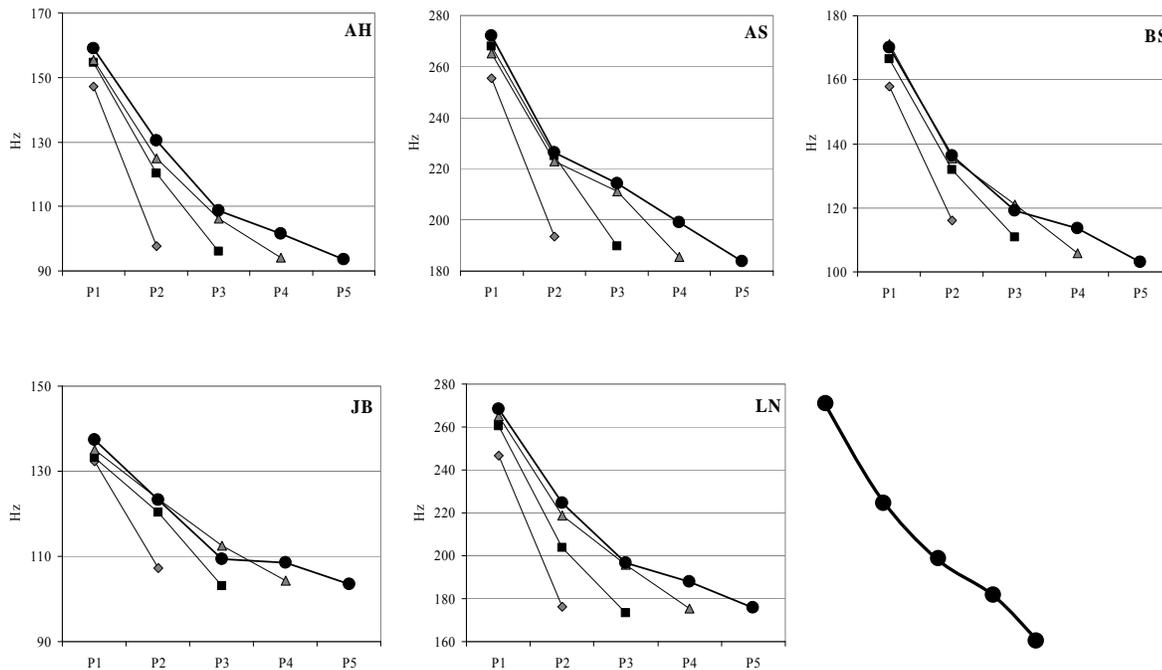

Fig. 4: Mean peak values for each of the MAE speakers, presented separately for each list length: diamonds are used for peaks in two-word lists, squares for peaks in three-word lists, triangles for peaks in four-word lists, and circles for peaks in five-word lists. Values for five-word lists adapted from Liberman and Pierrehumbert (1984) are presented in the bottom right corner for comparison.

The four- and five-word lists of all speakers except PM were modelled as exponential decay to a non-zero asymptote (see section 2.5). Results of the modelling for four-word lists are presented in Table 6 (the models and fit for both four- and five-word lists can be found in Appendix I). As can be seen from Table 6, not all data showed final lowering; indeed, while the data of four of the five SBE speakers showed lower final peaks than predicted by the models, only the data of AS among the MAE speakers followed the same pattern. Thus, only 50% of the overall data for four-word lists showed evidence for final lowering, and most of the cases were from the SBE group of speakers.

Table 6: Mean attested and predicted values (in Hz) for peaks in four-word lists, separately for each speaker; the top half of the table presents data from the SBE speakers (AB–SF), and the bottom half data from the MAE speakers (AH–LN). Cases in which the attested scaling of P4 was more than 3 Hz lower than predicted by the model are in bold.

	P1		P2		P3		P4	
Speaker	Attested	Predicted	Attested	Predicted	Attested	Predicted	Attested	Predicted
AB	193.8	193.8	177.5	177.5	174.1	174.1	**170.0**	**173.3**
DE	145.3	143.4	123.4	123.6	122.3	122.3	**118.8**	**122.2**
JR	143.4	121.1	97.2	97.2	90.3	90.3	88.7	88.2
RL	290.8	290.8	247.0	246.9	229.7	229.7	**214.3**	**222.9**
SF	265.2	265.2	228.8	228.8	219.3	219.3	**205.0**	**216.7**
AH	160.7	161.8	130.7	129.3	109.2	109.6	101.4	97.7
AS	271.9	271.9	226.2	226.2	214.2	214.2	**199.0**	**211.0**
BS	170.1	170.2	136.4	136.4	119.2	119.2	113.6	110.6
JB	137.0	138.0	123.2	121.7	109.4	109.9	108.3	101.5
LN	267.9	269.4	224.1	222.5	196.5	197.0	187.9	183.1

For five-words, different models were created for lists ending in *long beans* and *haricot beans*, since—as shown in section 3.2—the scaling of P5 was lower in the former. As can be seen in Table 7, which presents the differences between attested and predicted values in those models, five-word lists show more consistent evidence for final lowering than the four-word lists in both SBE and MAE. Since both subsets show final lowering (with a few exceptions), the data were pooled for presentation purposes in Figs 5 and 6. As can be seen by comparing the data in Table 7 and Fig. 5, three of the SBE speakers, JR, SF and RL, show strong evidence of final lowering; on the other hand, in DE's data, only the *haricot beans* subset shows final lowering, while in AB's data the predicted scaling of the final peak is lower than or almost the same as the attested scaling, and thus her data show no evidence for final lowering. In the MAE data, on the other hand, all speakers show evidence for final lowering in the *long beans* subset, but only two of them show final lowering in the *haricot beans* subset (see Fig. 6 and Table 7). Overall, the modelling of five-word lists showed evidence for final lowering in 70% of the data.

Table 7: Differences between attested mean F_0 values (in Hz) and predicted values for each peak in five-word lists; data are presented separately for lists ending in *haricot beans* and lists ending in *long beans*. The top part of the table presents results from the SBE speakers (AB–SF); the bottom half presents results from the MAE speakers (AH–LN). Cases in which the attested scaling of P5 was lower than predicted by the model by more than 3 Hz are in bold.

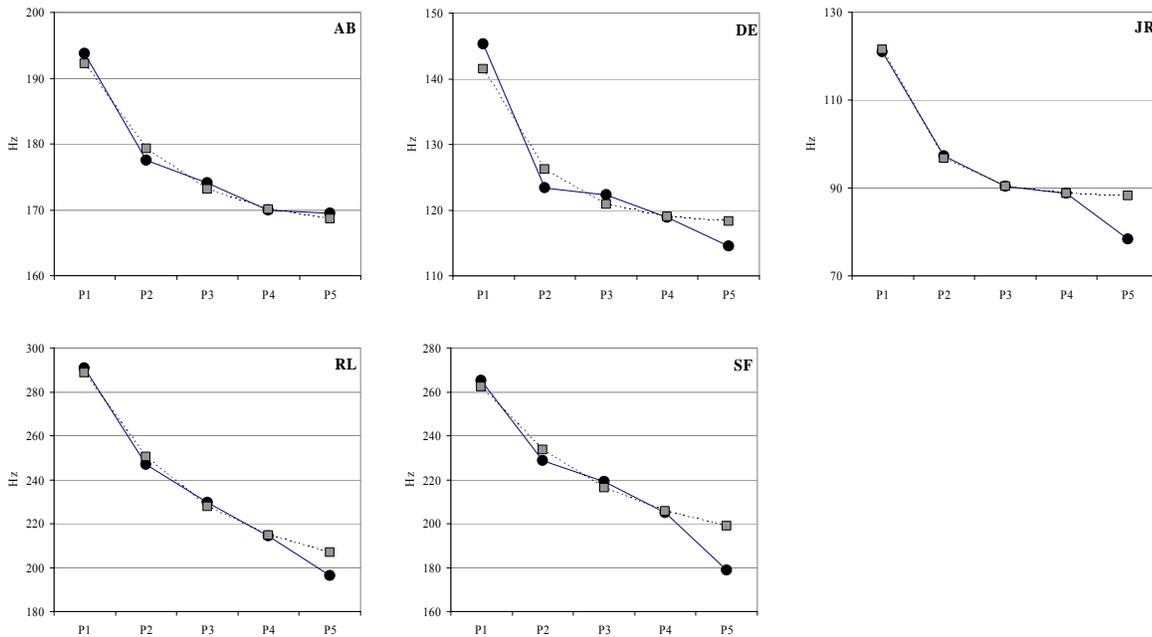|	P1 attested–predicted		P2 attested–predicted		P3 attested–predicted		P4 attested–predicted		P5 attested–predicted	
Speaker	*haricot*	*long*	*haricot*	*long*	*haricot*	*long*	*haricot*	*long*	*haricot*	*long*
AB	1.63	1.54	–2.03	–1.69	1.10	0.78	–0.15	–0.09	2.59	–0.76
DE	3.40	4.45	–2.79	–2.96	1.37	1.47	–0.13	–0.11	**–6.26**	–1.19
JR	0.66	–7.04	–0.42	2.30	0.07	–0.08	–0.00	–0.00	**–9.67**	**–10.42**
RL	1.95	2.41	–2.76	–3.87	1.40	2.28	–0.21	–0.40	**–6.60**	**–14.38**
SF	3.09	3.04	–4.99	–4.65	3.21	2.84	–0.57	–0.48	**–18.06**	**–22.15**
AH	–1.84	–2.64	2.30	3.42	–0.82	–1.18	0.09	0.13	–0.90	**–5.90**
AS	3.91	3.72	–5.03	–5.76	2.62	3.56	–0.37	–0.61	–1.28	**–17.18**
BS	–2.40	–0.85	2.40	0.85	–0.63	–0.25	0.05	0.02	**–5.89**	**–10.15**
JB	2.37	–4.40	0.61	4.10	–4.92	–0.67	2.80	0.01	4.41	**–4.61**
LN	–4.51	–2.63	4.72	3.03	–1.24	–0.97	0.10	0.10	**–3.47**	**–12.30**

Fig. 5: Mean attested and predicted F_0 values for peaks in five-word lists, separately for each SBE speaker; data are pooled across lists ending in *long beans* and *haricot beans*; for differences between the two lists, see Table 7. Black circles show attested values; grey squares show predicted values.
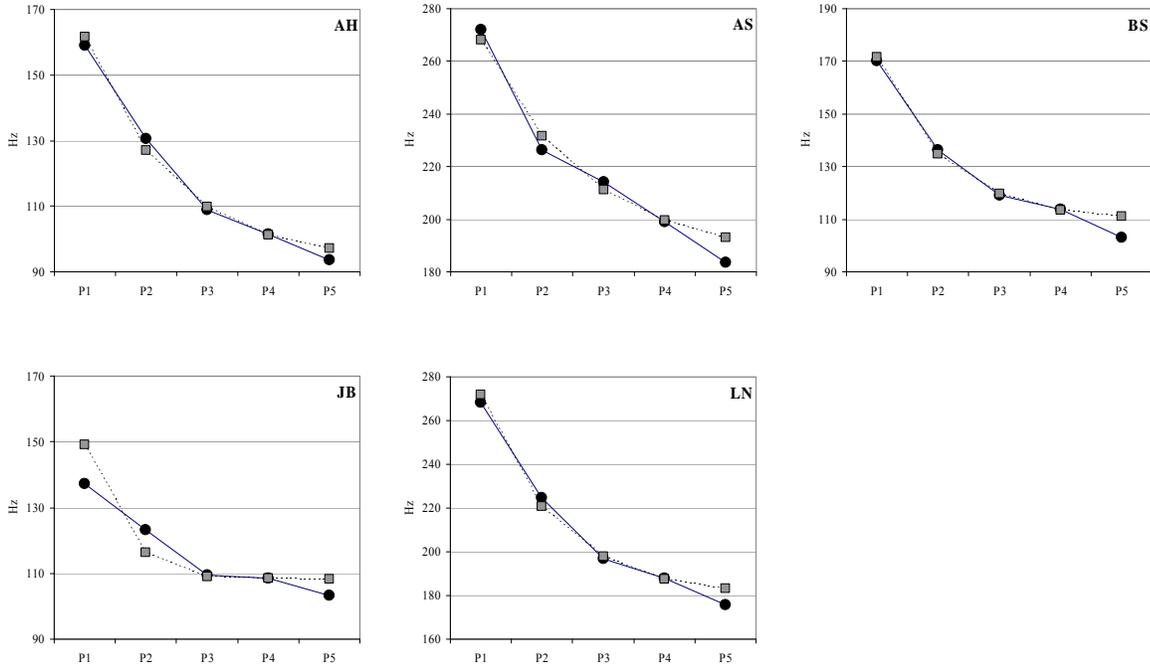
Fig. 6: Mean attested and predicted F_0 values for peaks in five-word lists, separately for each MAE speaker; data are pooled across lists ending in *long beans* and *haricot beans*; for differences between the two lists, see Table 7. Black circles show attested values; grey squares show predicted values.

3.5. Evidence for final lowering (iii): measuring interaccent steps

As mentioned, the final method used for final lowering detection involved comparing the steps in F_0 drop between successive peaks in three-, four- and five-word lists. These F_0 steps (in ERB) were subjected to ANOVAs with DIALECT as between subjects factor and STEP ORDER (first, second, third or fourth step) as repeated measures factor, in order to see if final steps would be larger than preceding steps, a result that would suggest that final lowering was present in the data.

The data for all three lists showed a main effect of STEP ORDER, but no effect of DIALECT [STEP ORDER for three-word lists, $F(1,8)=7.4$, $p<.05$; for four-word lists, $F(2,16)= 19.01$, $p<.01$; for five-word lists, $F(3,24)=43.95$, $p<.01$; DIALECT for three-word lists, $F(1,8)=1.1$, n.s.; for four-word lists, $F(1,8)= 1.12$, n.s.; for five-word lists, $F(1,8)=2.15$, n.s.]. As can be seen in Fig.

7, in three-word lists, the first step was larger than the second; in four-word lists, the first step was also larger than the next two steps [$p<.01$], but there was no difference between the second and third step. In five-word lists, the first step was also larger than the others [$p<.01$]; there were no differences between any of the other steps, although the difference between step 3 and step 4 came very close to significance [$F(1,8)=5.01$, $p<.055$], suggesting that the fourth (and final) step could indeed be larger than the third. (In five-word lists there was also interaction between DIALECT and STEP ORDER [$F(3,24)=4.3$, $p<.01$], which was due to the fact that in the MAE data step 2 was larger than step 3 [$p<.01$], a difference that was not present in the SBE data.) Although these statistical results suggest that final steps in F_0 were not larger than preceding steps, Fig. 7 and the data from five-word lists indicate that this null result is as likely to reflect a lack of statistical power due to the very small effect expected here, as it is likely to reflect a genuine absence of final lowering. If the former interpretation is correct, then this method, like the other two methods used here, provides evidence for final lowering in the present data, albeit weakly so.
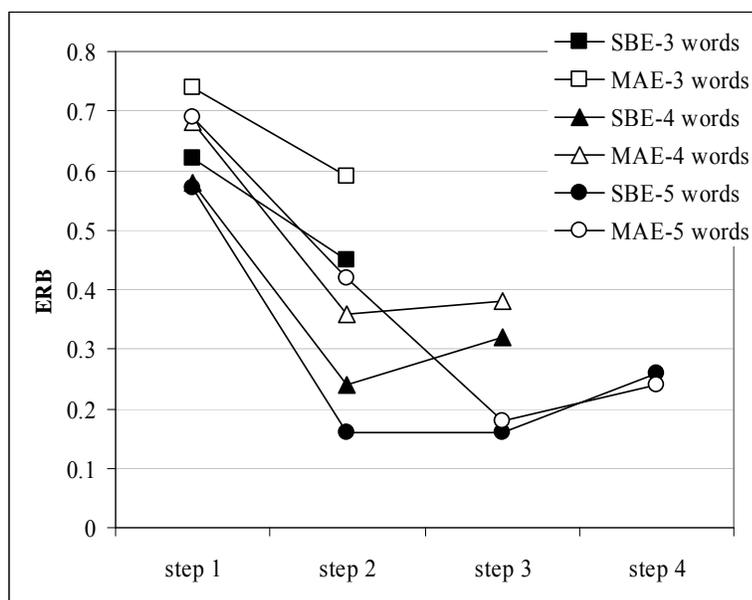

Fig. 7: Mean F_0 differences (in ERB) between successive peaks, for three-, four- and five-word lists; data pooled across last interaccent interval and final list item, but presented separately for each dialect (standard errors in these data ranged from 0.05 to 0.1).

4. Discussion

The results of this experiment showed that the extent to which final lowering can be said to be present in a set of data largely depends on the method used to detect its presence. Here, the comparison of peak scaling showed that final peaks are consistently lower than penultimate peaks (with the same order) in the data of both MAE and SBE speakers, suggesting a consistent presence of final lowering. Additionally, modelling the data using an exponential function showed lower scaling of attested than predicted peaks in 63% of the four- and five-word lists together. However, when the steps in F_0 drop between consecutive peaks were compared, there was no statistically significant evidence that the last step was larger than any preceding steps, a null result that has in the past been interpreted as evidence against the presence of final lowering (Grabe 1998). Taken together, these results suggest that final lowering has most probably a small effect on peak scaling (at least in lab speech), and for this reason it can be very sensitive to the method used for its detection. In particular, lack of statistical power may make the small differences in peak scaling that are to be expected when speakers are close to the bottom of their pitch range difficult to detect in an experiment with a small number of speakers who use different pitch ranges (even if the data are normalized, as in the present case). In this respect, a combination of modelling and comparison between final and penultimate peaks appears more robust at detecting small final lowering effects than the use of F_0 drops between successive peaks, which clearly requires a larger sample to show consistent and statistically significant results.

Despite the difficulties with measuring final lowering, the data do show that the scaling of final peaks is little affected by declination: although there is robust evidence that the duration of the last interaccent interval increases when it is lengthened by one syllable, this increase does not lead to a lowering of the final peak. Alternatively, we could say that the rate of declination is such that an increase of 27 ms does not result in a substantial difference in F_0 between final peaks preceded by two or three unstressed syllables. A crude model of declination for the present data, based on the drop in F_0 from the first to the final peak divided by the temporal distance between them, gives a declination rate of 48 Hz/s (1.82 ERB/s) or 1.3 Hz (0.05 ERB) during the 27 ms that the extra syllable adds to the last interaccent interval.[6] Certainly a difference of this magnitude would be difficult to detect statistically in a

small-scale experiment; on the other hand such a small effect would not have resulted in differences of the scale observed here and in Liberman and Pierrehumbert (1984). Thus, the hypothesis that final lowering is an artefact of Liberman and Pierrehumbert's experimental design is not supported by the present study. Rather, the data corroborate the conclusion of Liberman and Pierrehumbert (1984) that final lowering is independent of declination.

The alternative hypothesis advanced here, namely that SBE may not use final lowering in "listing" intonation, was weakly supported by the results: final lowering was not present in some SBE data, but this absence was concentrated on two particular speakers, AB who showed no evidence for final lowering in five-word lists, and PM who showed consistent declination reset due to the fact that he did not produce the longer lists as one phrase. This could indicate that there are speakers of SBE who do not employ final lowering regularly, at least with this tune. However, it must also be acknowledged that final lowering was regularly used by the majority of the SBE speakers, and that it was even more frequent than in MAE: the modelling showed final lowering in 73% of the SBE data, as opposed to 53% in the MAE data.

The extent to which the present results agree with those of previous studies depends on how final lowering is measured. If modelling and final vs. penultimate peak comparisons are used for detecting final lowering, the present data weakly agree with those of Grabe (1998) showing that for two of the six SBE speakers evidence for final lowering is absent or erratic. On the other hand, the present data agree better with Grabe's if her method of measuring final lowering is employed. As mentioned, the absence of statistically significant results when this method is used could simply reflect a small effect that was not detected due to lack of statistical power. This, however, does not necessarily mean that final lowering was present in Grabe's data and was simply not detected because of the methodology chosen. The reason for suggesting that final lowering may indeed have been absent from her data independently of measurement method, has to do with the fact that her materials though superficially similar to those used here and to those of Liberman and Pierrehumbert (1984) differ from them in one fundamental way. Specifically, although the materials of the present study—and those of Liberman and Pierrehumbert (1984)—have been referred to here as "word lists," they do in fact constitute legitimate utterances; e.g., as mentioned in section 2.3., they could be interpreted as answers to questions such as *Which kinds of beans do you like?*, or *What kinds of*

berries did you pick at the farm?. Grabe's materials, on the other hand, are closer to being word lists in the literal sense: they consisted of compounds with the same first morpheme, such as *moonlight, moonlit, moonbeam, moonshine, moonstone*, which are less likely to form a pragmatically coherent utterance. Lists like these could be interpreted only as answers to a metalinguistic question such as *Can you think of any words that begin with "moon"*; in such an answer, the test-words are not *used* but *mentioned* (Chris Barker, pers.com.) This difference in usage could well have elicited a different tune. This interpretation is supported by the pitch tracks presented in Grabe (1998) which do not show the "staircase" pattern, evident in both the MAE and SBE data here (see Fig. 1).[7] In other words, these small differences in materials may have resulted in large differences in how the speakers interpreted their task in each experiment, and consequently their use of final lowering. If the lists were not seen as plausible discourse units and if final lowering is a cue to discourse finality in SBE as it is in MAE (Pierrehumbert and Hirschberg 1990; Herman 2000), then it is possible that Grabe's subjects simply did not use final lowering because it was not appropriate for the task they thought they were performing.

The differences between the present study and Liberman and Pierrehumbert (1984) are more a question of degree, in that the evidence for final lowering is not as consistent here as in their study. One possible reason for the difference could be that, as Nolan (1995) points out, Liberman and Pierrehumbert used as subjects themselves and a third speaker to whom the expected tune was illustrated by example and whose ability to produce it was checked *before* the experiment (Liberman and Pierrehumbert 1984: 172). In contrast, the present speakers were not trained phoneticians and were not given any instructions as to how to produce the materials. This does not necessarily suggest that Liberman and Pierrehumbert's results were artificial, but could mean that their speakers had more consistently construed each utterance as being discourse final because of their training, while the present speakers did not. This difference may also be the reason why in the four-word utterances most MAE speakers did not show final lowering, while the SBE speakers did: the MAE speaking rate was faster than SBE, and this could have resulted in the MAE speakers "running" their shorter utterances together, and therefore not treating them as discourse final [in an ANOVA with DIALECT as a between subjects factor and utterance length as the dependent variable, SBE utterances were found to be 230 ms longer on average; F(1,9)=6.13, *p*<.05].

Finally, the results showed that the position of the last accent early or late with respect to the end of the utterance (*haricot beans* and *long beans* subsets respectively) did not affect final peak scaling except in five-word lists, in which the peak of *haricot beans* was scaled higher than that of *long beans*. At first glance, this difference could be seen as evidence that final lowering is "gradient and temporally diffuse" (Gussenhoven 2004: 121), and thus a purely physiological aspect of the phonetic realization of tonal scaling. Yet, as shown in 3.3., the peaks of both the *haricot beans* and the *long beans* subset showed lower than predicted scaling (even though the result was more robust for *long beans*, as shown in Table 7). Therefore, the observed differences between the *haricot beans* and *long beans* sets cannot be straightforwardly attributed to a physiological mechanism that does not stretch as far back as the peak of *haricot beans*, as was originally hypothesized.

A more likely interpretation rests on the operation of downstep and its interaction with the term *haricot beans* which was unfamiliar to many MAE speakers. A look at the data of Table 7 shows that final lowering was not present in the *haricot beans* subset of three out of five MAE speakers who, as a group, were less comfortable with the term *haricot beans* than the SBE speakers. This unfamiliarity could have prompted them to suspend downstep on *haricot beans*; this is also indicated by the fact that in the MAE data the differences between attested and expected peak scaling were generally smaller for the *haricot beans* subset (see Table 7). The fact that the same effect is not observed in the shorter lists could suggest that final lowering is less pronounced in shorter utterances (e.g., Liberman and Pierrehumbert 1984: 222); if so, then statistical differences between peaks on *haricot beans* and *long beans* may not emerge in short lists. To put it differently, in the shorter lists the effect of final lowering on the peak of *long beans* was small and thus the scaling of those peaks was statistically indistinguishable from the lack of downstep on *haricot beans*. In the five-word lists, the lack of downstep on *haricot beans* made their peaks stand out relative to the peaks of *long beans*, which had both downstep and substantial final lowering due to the longer utterance length. Taken all together, these results suggest that more research is needed not only on the scope of final lowering but also on the extent to which it operates in utterances of different function and duration.

5. Conclusion

In conclusion, the present experiment showed evidence for final lowering in both SBE and MAE, but the extent to which its presence was detected depended on the method used: data modelling and final vs. penultimate peak comparisons turned out to be more sensitive and thus more appropriate for detecting final lowering than the comparison of drops in F_0 between successive peaks. In addition, the results showed that the scaling of final peaks was not influenced by declination, supporting the view that final lowering is an independent phenomenon. Further, although final lowering was present in both SBE and MAE, its application was not consistent; it was suggested that this variability could be related to the small magnitude of the effect, which does not allow for easy detection, but also to the extent to which the speakers can consider their utterances in a laboratory speech task to be discourse final. This possible interpretation, coupled with the fact that final lowering targeted the last accent rather than an unspecified stretch at the end of the utterance, indicates that final lowering is grammaticalized in the linguistic varieties examined here.

Appendix I

Best-fit exponential models for each speaker; models for four-word lists and separately for five-word lists ending in *haricot beans* and lists ending in *long beans* are presented. Goodness of fit is indicated by the correlation coefficient to the right of each model. Models for the SBE speakers are presented at the top (AB–SF); models for the MAE speakers are found at the bottom of the table (AH–LN).

Speaker	Models	c.c.
	$AB_{4words} = 98.601 * (0.209)^{1-5} + 173.151$	–0.999
AB	$AB_{haricot} = 47.838 * (0.514)^{1-5} + 166.498$	–0.986
	$AB_{long} = 55.064 * (0.457)^{1-5} + 167.969$	–0.991
	$DE_{4words} = 322.411 * (0.066)^{1-5} + 122.182$	–0.998
DE	$DE_{haricot} = 62.913 * (0.381)^{1-5} + 117.81$	–0.972
	$DE_{long} = 70.349 * (0.327)^{1-5} + 117.972$	–0.96
	$JR_{4words} = 114.633 * (0.293)^{1-5} + 87.374$	–0.999
JR	$JR_{haricot} = 100.026 * (0.331)^{1-5} + 86.738$	–0.999
	$JR_{long} = 282.904 * (0.139)^{1-5} + 89.436$	–0.995
	$RL_{4words} = 183.932 * (0.393)^{1-5} + 218.473$	–0.999
RL	$RL_{haricot} = 160.753 * (0.556)^{1-5} + 198.234$	–0.998
	$RL_{long} = 156.794 * (0.616)^{1-5} + 193.083$	–0.996
	$SF_{4words} = 188.116 * (0.263)^{1-5} + 215.856$	–0.999
SF	$SF_{haricot} = 121.261 * (0.625)^{1-5} + 186.235$	–0.988
	$SF_{long} = 118.769 * (0.599)^{1-5} + 191.022$	–0.989
	$AH_{4words} = 136.196 * (0.607)^{1-5} + 79.2$	–0.999
AH	$AH_{haricot} = 149.15 * (0.498)^{1-5} + 91.555$	–0.998
	$AH_{long} = 130.67 * (0.513)^{1-5} + 92.71$	–0.996
	$AS_{4words} = 235.4 * (0.263)^{1-5} + 209.874$	–0.999
AS	$AS_{haricot} = 150.681 * (0.524)^{1-5} + 186.969$	–0.992
	$AS_{long} = 147.649 * (0.603)^{1-5} + 181.159$	–0.989
	$BS_{4words} = 135.151 * (0.507)^{1-5} + 101.645$	–0.999
BS	$BS_{haricot} = 154.075 * (0.412)^{1-5} + 109.513$	–0.998
	$BS_{long} = 152.695 * (0.402)^{1-5} + 109.163$	–0.999
	$JB_{4words} = 81.187 * (0.722)^{1-5} + 79.408$	–0.996
JB	$JB_{haricot} = 70.755 * (0.790)^{1-5} + 78.183$	–0.943
	$JB_{long} = 96.443 * (0.369)^{1-5} + 106.216$	–0.988
	$LN_{4words} = 188.985 * (0.544)^{1-5} + 166.54$	–0.999
LN	$LN_{haricot} = 224.978 * (0.423)^{1-5} + 179.444$	–0.997
	$LN_{long} = 191.467 * (0.458)^{1-5} + 180.566$	–0.998

Notes

[1] This aspect of the experimental design could have a small effect on the comparison between penultimate and final peaks, since the vowels of penultimate peaks (the stressed vowels of *navy* and *green*) are of greater height than the vowels of final peaks (the stressed vowels of *long* and *haricot*). However, any intrinsic F_0 effects are likely to be negligible in these peaks for several reasons. First, the average intrinsic F_0 effect of 15.3 Hz reported by Whalen and Levitt (1995) pertains to [i] and [u] vs. [a] when these vowels are uttered in isolation, or as part of test-words (e.g., *heed*) elicited either in isolation or within short carrier phrases. Differences are much smaller when running text is used, especially towards the end of utterances where F_0 tends to be lower (e.g., Ladd and Silverman, 1984; Shadle 1985; Steele 1986); similarly small effects apply to vowels with low tone in tone languages (Connell 2002). As Whalen and Levitt (1995: 362) succinctly put it, "in the low region, IF_0 disappears." This lack of intrinsic F_0 effects when F_0 is low is probably related to the finding that the magnitude of intrinsic F_0 correlates positively with pitch range (Reinholt Petersen 1978; Connell 2002). For the present data this means that any intrinsic F_0 effects are likely to be even smaller, at least for speakers AB, DE, JR and JB, who have very narrow pitch ranges (see Table 3). In addition, although intrinsic F_0 effects have been clearly demonstrated for high vowels [i] and [u] vs. the low vowel [a], the intrinsic F_0 of mid-vowels has been shown to be intermediate between that of high and low vowels (Reinholt Petersen 1978). In the present data, three out of the four test-words in question have—at least in some data—high-mid or low-mid vowels in stressed position: *navy* ([e] and [eɪ] for MAE and SBE speakers respectively), *haricot* ([æ], but also [ɛ] for some speakers), and *long* ([ɒ] for SBE speakers, and [ɔ] for the MAE speakers who have the [ɑ] : [ɔ] contrast). Finally, it should be noted that the effect of intrinsic F_0 on perception is very small: in paired comparisons listeners increase the F_0 of high vowels [i] and [u] by 1.25-2.8 Hz in order to hear them as having the same pitch as the low vowel [a] (for a review see Fischer-Jørgensen 1990: 101). The above observations suggest that F_0 differences of the magnitude reported here are unlikely to be due to intrinsic F_0 effects in this position in the utterance, though it is conceivable that such effects have added to the magnitude of the final lowering results; in any case, the differences found here are too large not to result in the perception of the final peaks as being lower in pitch than the penultimate peaks and lower than otherwise expected.

[2] The autosegmental analysis of this tune is not entirely clear: Liberman and Pierrehumbert (1984) refer to the accent of this tune as "a pair of tones, H+L" (p. 217), while Beckman and Pierrehumbert (1986) say that these tunes consist of sequences of "H*+L or H+L* accents" (p. 280). Since Beckman and Pierrehumbert (1986) also describe H*+L as an accent which is realized with a peak on the stressed syllable, while H+L* is described as having a peak before the stressed syllable, H*+L seems a more plausible representation of the accents in the present data. For an alternative analysis, see Gussenhoven (2004: 122) and references therein.

[3] The late alignment of the first peak in the British data was probably due to dialectal differences in tonal realization rather than a difference in pitch accent (cf. Arvaniti and Garding in press, Atterer and Ladd 2004).

[4] Rossi (1971) and Klatt (1973) are among the few studies that relied on speech-like stimuli; their results differ considerably from each other but, as 't Hart et al. (1990: 31) point out, Rossi's thresholds are "exceptionally high;" this is why 3 Hz, which is a rounded-up version of Klatt's highest limen (2.5 Hz), was adopted here.

[5]The ANOVAs presented in sections 3.2 and 3.3 were also run without PM's data but the results were not different from those presented here in which his data were included.

[6] A model of this sort assumes that changes in the F_0 topline are due exclusively to declination, and this is

certainly not the case, as the modelling of the present data as exponential decay shows. Although the issue of the role of declination in tone scaling is by no means resolved, a detailed discussion of it is beyond the scope of this paper.

[7] Although the accents in Grabe (1998) are transcribed as H*+L, they are not F_0 drops from one level to the next, as in Liberman and Pierrehumbert (1984); sustained F_0 levels, were, however, present in the data of some of her speakers, whose accents were transcribed as H*> (Grabe 1998: 208). A detailed discussion of these differences and what they entail is beyond the scope of this paper.

References

Arvaniti, Amalia and Garding, Gina   in press Dialectal variation in the rising accents of American English. In: Jennifer Cole and José Hualde (eds.), *Papers in Laboratory Phonology 9*. Mouton de Gruyter.

Atterer, Michaela and Ladd, D. Robert 2004   On the phonetics and phonology of "segmental anchoring" of F_0: evidence from German. *Journal of Phonetics* 32: 177–197.

Beckman, Mary E.   1986   *Stress and Non-Stress Accent*. Dordrecht: Foris.

Beckman, Mary E. and Pierrehumbert, Janet B. 1986   Intonational structure in Japanese and English. *Phonology Yearbook* 3: 255–309

Connell, Bruce   2002   Tone languages and the universality of intrinsic F_0: evidence from Africa. *Journal of Phonetics* 30: 101–129.

Connell, Bruce and Ladd, D. Robert   1990   Aspects of pitch realization in Yoruba. *Phonology* 7: 1–30.

Cooper, William E. and Sorensen, John M. 1981   *Fundamental frequency in sentence production*. New York: Springer-Verlag.

Dalton, Martha and Ní Chasaide, Ailbhe   this volume   Tonal alignment in Irish dialects.

Fischer-Jørgensen, Eli   1990   Intrinsic F_0 in tense and lax vowels with special reference to German. *Phonetica* 47: 99–140.

Grabe, Esther   1998   Intonational phonology: English and German. Ph.D. dissertation, Max-Planck Institut for Psycholinguistics and University of Nijmegen.

Grabe, Esther, Post, Brechtje, Nolan, Francis and Farrar, Kimberley   2000   Pitch accent realization in four varieties of British English. J*ournal of Phonetics* 28: 161–185.

Gussenhoven, Carlos 2004　*The Phonology of Tone and Intonation.* Cambridge: Cambridge University Press.

Gussenhoven, Carlos and Rietveld, A. C. M.　1988　Fundamental frequency declination in Dutch: Testing three hypotheses. *Journal of Phonetics* 16: 355–369.

't Hart, Johan, Collier, René and Cohen, Antonie 1990　*A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody.* Cambridge: Cambridge University Press.

Herman, Rebecca　1996　Final lowering in Kipare. *Phonology* 13: 171–196.

Herman, Rebecca　2000　Phonetic markers of global discourse structures in English. *Journal of Phonetics* 28: 466–493.

Herman, Rebecca, Beckman, Mary and Honda, Kiyoshi　1996　Subglottal pressure and final lowering in English. In: H. Timothy Bunnell and William Idsardi (eds.), *Proceedings of the 4th International Conference on Spoken Language Processing*, 145–148. Philadelphia.

Hermes, Dik J. and van Gestel, Joost C.　1991　The frequency scale of speech intonation. *Journal of the Acoustical Society of America* 90: 97-103.

Klatt, Dennis H. 1973　Discrimination of fundamental frequency contours in synthetic speech: implications for models of speech perception. *Journal of the Acoustical Society of America* 53: 8–16.

Labov, William, Ash, Sharon and Boberg, Charles　in press　*Atlas of North American English.* Mouton de Gruyter. (http://www.ling.upenn.edu/phonoatlas/Ch11_2nd.pdf)

Ladd, D. Robert 1988　Declination: a review and some hypotheses. *Phonology Yearbook* 1: 53–74.

Ladd, D. Robert and Silverman, Kim E. A.　1984　Vowel intrinsic pitch in connected speech. *Phonetica* 41: 31–40.

Laniran, Yetunde O.　1992　Intonation in tone languages: the phonetic implementation of tones in Yoruba. Ph.D. Dissertation, Department of Linguistics, Cornell University.

Liberman, Mark and Pierrehumbert, Janet　1984　Intonational invariance under changes in pitch range and length. In: Mark Aronoff and Richard T. Oehrle (eds.), *Language Sound Structure*, 157–233. Cambridge MA: The MIT Press.

Max, Ludo and Onghena, Patrick   1999   Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research* 42: 261–270.

Nolan, Francis   1995   The effect of emphasis on declination in English intonation. In: Jack Windsor Lewis (ed.), *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor*, 241–255. London: Routledge.

O'Shaughnessy, Douglas and Allen, Jonathan   1983   Linguistic modality effects on fundamental frequency in speech. *Journal of the Acoustical Society of America* 74: 1155–1171.

Pierrehumbert, Janet 1980   The phonology and phonetics of English intonation. Ph.D. dissertation, MIT.

Pierrehumbert, Janet and Beckman, Mary   1988   *Japanese Tone Structure*. Cambridge, MA: The MIT Press.

Pierrehumbert, Janet and Hirschberg, Julia 1990   The meaning of intonational contours in the interpretation of discourse. In: Philip R. Cohen, Jerry Morgan and Martha E. Pollack (eds.), *Intentions in Communication*, 271–311. Cambridge, MA: The MIT Press.

Poser, William   1984   The phonetics and phonology of tone and intonation in Japanese. Ph.D. Dissertation, MIT.

Prieto, Pilar, Shih, Chilin and Nibert, Holy 1996   Pitch downtrend in Spanish. *Journal of Phonetics* 24: 445–473.

Prieto, Prieto, D'Imperio, MariaPaola and Gili Fivela, Barbara   in press Pitch   accent alignment in Romance: primary and secondary associations with metrical structure. *Language and Speech*.

Reinholt Petersen, Niels   1978   Intrinsic fundamental frequency of Danish vowels. *Journal of Phonetics* 6: 177–189.

Rossi, Mario 1978   La perception des glissandos descendants dans les contours prosodiques. *Phonetica* 35: 11–40.

Shadle, Christine H.   1985   Intrinsic fundamental frequency of vowels in sentence context. *Journal of the Acoustical Society of America* 78: 1562–1567.

Steele, Shirley A.1986   Interaction of vowel F_0 and prosody. *Phonetica* 43: 92–105.

Thorsen, Nina   1985   Intonation and text in Standard Danish. *Journal of the Acoustical Society of America* 77: 1205–1216.

Trudgill, Peter   1999   *The dialects of England*. Oxford: Blackwell.

van den Berg, Rob, Gussenhoven, Carlos and Rietveld, Toni 1992   Downstep in Dutch: Implications for a model. In: Gerard J. Docherty and D. Robert Ladd (eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 335–359. Cambridge: Cambridge University Press.

Whalen, Douglas H. and Levitt, Andrea G. 1995   The universality of intrinsic F_0 of vowels. *Journal of Phonetics* 23: 349–366.