

Durational prosody and topic organization: differences between English and French

Caroline L. Smith

caroline@unm.edu

University of New Mexico

Summary: This study focuses on discourse-level prosody by examining durational patterns that relate to the type of topic transition between sentences, that is, the relation between the topic of one sentence and the topic of the next. The study compares these patterns in texts read aloud in English and French. The topic organization of each text was analyzed by native speakers who categorized each sentence-to-sentence transition as a Topic Shift, Continuation, Elaboration or Text Marker. Recordings of three English and two French speakers show effects of Topic Transition Type on durations in both languages. The effects differ, but are of similar magnitude. Next, one recording from each language was chosen to use in a modeling procedure which manipulated rate, sentence-final lengthening and pause duration to create multiple versions of the original recording. The different versions were presented to listeners of appropriate language background. Of the manipulated versions, English listeners preferred those where the durations matched the speakers' means for each separate type of topic transition. French listeners preferred the version where each prosodic variable kept the same mean value throughout. These results suggest that topic-related durational patterns occur in both languages but may be more salient in English.

1. Introduction

In a number of studies, discourse- or text-level organization has been found to correlate with variation in the prosodic dimensions of pitch, duration and acoustic energy (e.g., Lehiste 1979, Grosz & Hirschberg 1992, Ayers 1994, Swerts & Geluykens 1994, Hirschberg & Nakatani 1996, Noordman et al. 1999, van Donzel 1999, Herman 2000, Wichmann 2000). It is sometimes assumed that speakers may be motivated to produce this variation, at least in part, as a means of conveying the structure of their message to listeners. This kind of variation is found in speech produced spontaneously as well as in pre-planned speech such as when a familiar text is read aloud. The latter is the style of speech studied here. Although this kind of variation is well-documented in production, very little is known about listeners' reactions to it. Does speakers' production of discourse-level variation have any effect on listeners? Here we focus on durational variation, and investigate listeners' sensitivity to variation that is related to the organization of a text being read aloud.

The particular focus of this study is cross-linguistic, comparing English and French. These languages are known to have very different lexical and phrasal prosody, so it would not be surprising if they also exhibit different prosodic patterns as a function of the organization of longer stretches of spoken material (Vaissière 1997). The traditional labeling of English and French as 'stress-' and 'syllable-timed', respectively, may be misleading and over-simplified (Dauer 1983, Bertinetto 1989), but it does indicate that the languages differ as to how temporal patterning is established. A metric for acoustic rhythmic differences was proposed by Ramus, Nespors and Mehler (1999). They found that French had a greater proportion of vocalic intervals and a lower variance in the duration of consonantal intervals than English, and predict that listeners can use these dimensions to categorize these and other languages rhythmically. This proposal is based on differences in languages' phonological structures, and thus differs from other descriptions of cross-linguistic rhythmic differences that are based on durational patterns in production, such as differences in the amount of phrase- or sentence-final lengthening found in English, French and Swedish (Fant, Kruckenberg and Nord 1991).

Although English and French are often referred to as having different rhythms, an extensive comparison of durational patterns produced in English and French radio interviews (Grosjean and Deschamps 1975) found similar distributions of values in the two languages for such variables as speech rate and pause duration. Although the languages did differ significantly in several variables, the overall picture they give is of a broadly similar use of time when measured globally, including for example, similar proportions of phonation and pause time during speech. Similarly, the picture painted by Vaissière (1983) is of broadly similar use of prosodic dimensions, but different “priorities” in English and French. These results suggest that these languages may indeed differ in their durational properties, but that it is possible the differences may be localized, as in different amounts of lengthening, rather than global.

The study here focuses on durational patterns in relation to an entire discourse, here a text read aloud, but investigates them in quite local features, including changes in duration of the syllable and word immediately adjacent to a sentence boundary, as well as differences in speaking rate in the vicinity of a boundary. By comparing these patterns in English and French, we add to the very limited literature of cross-linguistic studies on discourse prosody. A notable contribution to this literature is Fon (2002) who compared English, Japanese, and Putonghua and Guoyu Mandarin, focusing on local prosodic differences at different types of discourse boundaries by examining patterns of duration and F0 within two syllables of a boundary. The present study investigates some of the same variables as Fon (2002), but differs from her study in that here the production data is a starting point for an additional angle of approach, that of listeners’ attitudes towards the discourse-related durational patterns.

Besides the interest in cross-linguistic comparison, the specific approach taken here was chosen to be potentially useful for implementation in synthesis systems. A major challenge in synthesis remains the production of appropriate, non-repetitive prosody during longer stretches of speech. This study investigates the idea of achieving this by adapting the prosody to the organization of topics in the material to be spoken. In addition, the timing models for synthesis, although complex, are still largely deterministic, with the same output produced each time that a particular combination of contextual factors occurs. Sentence and paragraph-level factors are less well understood than segmental and lexical factors, with the result that synthesis models may include fewer choices for factors with larger scope, and apply the same values to many contexts. Another issue is that even taking into account a wide range of factors, there is greater variability in natural speech than can be accounted for by current models. Here we investigate the possibility of compensating for this by modeling variability as well as the effects of factors related to topic organization.

In order to maximize the similarity with the task of a speech synthesizer, the style of speech chosen for study here is reading aloud. Synthesizers, even those in a dialogue system, tend to use a speaking style closer to what a human uses in reading a prepared text than in conversational speech. The texts selected for the current study were extracts from software manuals, chosen to be representative of the type of material that might be part of a computer help system. Another advantage of using this kind of text is that it has a well-defined, usually linear sequencing of topics, which facilitates the task of describing its topical structure. The method used to analyze the organization of topics in the texts was simple, with the intention that it be possible to automate it as part of a synthesizer. This would make it possible to automatically generate prosodic effects of topic organization.

The study described here differs from many others in examining both the speakers’ productions of durational patterns and listeners’ reactions to them. Most research into discourse-level patterns of prosody has focused on speakers’ behavior, but a few studies have looked at the relation between speakers’ productions and listeners’ ability to identify discourse boundary locations, or the presence vs., absence of discourse boundaries, in speech samples (e.g., Kreiman 1982, Swerts & Geluykens 1994, Geluykens & Swerts 1994). Fon (2004) found that although the number of cues marking a boundary seems to play a role in listeners’ ability to perceive it, so do the listeners’ expectations and preparedness. However, none of the studies cited here have explicitly addressed the importance of discourse boundary marking to listeners. If the goal is to improve synthetic speech, the most important question about discourse boundary marking is whether listeners want, or expect, to hear it. If they do, then synthesized speech should be constructed to include this type of pattern. If listeners do not wish to hear such patterns, they can safely be

omitted. The present study begins to investigate this question, with the added goal of determining whether English and French listeners have similar or different preferences.

2. Production study

The production study examined speakers' productions of transitions from one sentence to the next while reading a prepared text. The goal is to identify which acoustic durational properties distinguish transitions between pairs of sentences that are more or less closely related in topic. The durational properties studied here were lengthening of sentence-final words, duration of the pause between sentences, and speech rate preceding and following the sentence boundary. The study was conducted with speakers of English and French. The results of this study are then used as the basis for modeling and a study of listener preferences, reported in section 3.

2.1. Method

2.1.1 Texts and control sentences

The texts chosen for study were extracts from computer manuals. For English, a selection about spell-checking was extracted from the manual for the drawing program Canvas, and for French, a passage on creating a presentation was selected from a beginner's guide to PowerPoint. The English text was 60 sentences long; the French text was 79 sentences long. Thus the number of transitions between sentences within the text was 59 and 78 respectively.

One of the durational parameters studied here was lengthening of the sentence-final words. This was calculated by comparing the duration of the 'target' words as they occurred sentence-finally in the text, with their duration in a non-sentence-final context. The non-final context was constructed by creating control sentences so that each target word occurred sentence-medially in a control sentence. The length of the control sentences was controlled so that there were three syllables before the target word and eight syllables after it. (Since the target words themselves varied between one and five syllables in length, the total number of syllables in the control sentences varied accordingly.) Final lengthening was not calculated for list numbers, which occurred in both languages' texts.

In determining the amount of lengthening, it was necessary to control for prosodic prominence, because words that are prosodically prominent tend to be longer in both English (Turk and White 1999) and French. In order to control for this effect in English, the target words in all readings of the text and the control sentences were categorized as accented or unaccented. The experimenter and an assistant listened to the recordings and decided, for each reading, whether the target word was produced with a pitch accent or not. Difficult cases were resolved by consensus, but for four readings of words in the text passage, no decision could be reached so they were excluded from further analysis. The control sentences were designed to favor the production of the target word with or without a pitch accent depending on whether the presence or absence of accent seemed likely for that word in the text passage. For words which occurred in two or more accentually different contexts, or where it was unclear whether they were likely to be accented, two control sentences were constructed, one intended to favor an accented production and one favoring an unaccented production. There were ten words for which two control sentences were used, making a total of 49 control sentences.

For French, there were a total of 51 control sentences, with one for each word that occurred sentence-finally in the text. The structure of the control sentences placed the target word in a sentence-medial position where it was usually the final word of an accent group, a prosodically prominent position. (Here I use the term 'accent group' for the basic phrasing unit in French, which is known under various names [see, among many others, Hirst & Di Cristo 1984, Vaissière 1991, Lacheret-Dujour & Beaugendre 1999, and Jun & Fougeron 2000 for a slightly different definition].) The prominence of the target word was verified by listening for the pitch rise which normally occurs at the end of a non-final accent group (e.g., Di Cristo & Hirst 1997, Di Cristo 1998). Being final in the accent group meant that the target words were subject to lengthening, as is characteristic of this prosodic position (e.g., Padeloup 1992, Delais 1994, Post 2000). However, in one control sentence neither speaker made the target word prominent, and

in another, speaker F1 did not. These, as well as the list numbers, were excluded from calculations of final lengthening. The fact that the target words were somewhat lengthened in the control sentences means that their position sentence-finally in the text is likely to result in less additional duration than would be found in comparison to a non-prominent sentence-medial position, giving a lower estimate of final lengthening (which seems to be more speaker-specific and rate-dependent in French (Duez 1999, Astésano 2001) than in English). But this is probably unavoidable given the method being used. Almost all the target words were content words, which are likely to be produced as accent group-final, therefore prominent, whether sentence-medial or final. Final lengthening in French.

2.1.2 Text analysis

As described above, the texts used as the basis for the study were extracts from software manuals. The goal was to have an analysis of the topic organization of the text independent of any specific production, so that prosodic correlates of the topic organization can be tested for without risk of the circularity that can arise if discourse structure is analyzed on the basis of an audio recording (Swerts 1997). Thus the analysis of topic organization was based on an unformatted written text. The method of labeling the topic organization was adapted from that used by Nakajima and Allen (1993). Because their original scheme was devised for the analysis of spontaneous speech, some modifications were necessary in order to apply it to the instructional texts used here. The analysis describes the structure of a discourse in terms of the type of transition between successive conversational turns, or as applied here to written texts, between successive sentences. The four categories used in this study are shown in Table 1. Note that the label for each sentence-to-sentence transition was associated with the sentence preceding the transition.

Transition Type	Characteristics
Topic Shift	The next sentence initiates a new topic or substantially changes the topic under discussion.
Topic Continuation	The next sentence continues the same topic, providing new information which advances the progression of instructions.
Elaboration	The next sentence provides additional detail about the topic of the previous sentence.
Text Marker	Analogous to a discourse marker in spoken discourse, these are overt indications of textual structure, such as a list number.

Table 1: Topic Transition Types used for labeling the organization of the texts

It is also desirable for the analyses of the English and French texts to be as similar as possible, which would not be feasible if the analysis could only be performed by persons with extensive training in a particular theory. This consideration was another motivation for using a simple labeling scheme. It is also the case that previous studies (Noordman et al. 1999, Wightman et al. 1992) have found distinct acoustic correlates for only a limited number of distinct levels in a discourse structure hierarchy, suggesting that labeling schemes with many categories may be overly complex for this type of study. Rather than calling on one or two trained analysts, five American and four native French-speaking linguists were asked to analyze the texts. Linguists were chosen because, although they had no prior familiarity with the particular labeling scheme, it was believed that they would find the task easier than naïve language users. The assumption is that the native-speaker linguists' perception of the topic organization of the text will be similar to that of the speakers who recorded the texts. This assumption is critical since it is being assumed that the durational patterns produced by the speakers are meaningfully related to their understanding of the topic organization of the text.

2.1.3 Speakers and recordings

The data presented here come from recordings of three speakers of American English and two of Parisian French. All were young adults, aged from mid 20's to early 30's. The American English speakers will be referred to as E1, E2 and E3. Speaker E1 was male, the others female. The French speakers will be referred to as F1 (male) and F2 (female). Recordings were made in a quiet room using a Sony Professional Walkman with a Shure head-mounted microphone. A total of ten recordings were made of each speaker over a period of months, with mean intervals between recordings of 7 – 11 days. In addition to the materials discussed here, an additional text with matching control sentences was also recorded at the same time. At each recording session, the speakers were presented with a different order of these four sections (two texts and two sets of control sentences). Within each set of control sentences, the order of the sentences was randomized differently for each session, and two filler sentences were recorded at the beginning and end of each of the sets of control sentences. The recordings were digitized at a 10kHz sampling rate on a Kay Elemetrics CSL system. Durational measurements were made using the waveform and spectrograms on CSL, except for speaker F2, whose data were measured using Praat.

2.1.4 Measurements

Measurements were made of three durational parameters: pause duration, final lengthening and speech rate. The methods for each are described here.

Pause duration. The duration of the pause was measured between each pair of sentences. The release burst following a sentence-final stop was included in the duration of the pause. All pauses which coincided with page turns were excluded from analysis; there were two of these in each reading of each text, since both texts covered three pages.

Final lengthening. Final lengthening was determined, as described above, by comparing the acoustic duration of words which occurred sentence-finally in the text, with their duration when sentence-medial in a control sentence. Readings of control sentences were excluded if the speaker paused for more than 150 ms in the middle of a sentence, which occurred only with speaker F1 (15 readings excluded). In addition to measuring the duration of the entire target word, in English, the duration of the rime of the final syllable of the target word was also measured, because this unit has previously been found to be where the most lengthening takes place (Wightman et al. 1992). The rime was defined as the vowel and any following consonants. A syllabic consonant was also treated as a rime. In French, the duration of the final syllable having a full (non-schwa) vowel was measured, as the syllable seems to be the primary unit over which timing is regulated (Astésano 2001).

For the control sentences, the mean durations of the ten repetitions of each target word, and their rimes, and the corresponding standard deviations, were calculated. These will be referred to as μ -s and σ -s, respectively. Using a technique similar to that used by Wightman et al. (1992), the duration of each target word in a text sentence was pseudo-normalized. If d-t is the duration of one reading of a target word in the text, then the pseudo-normalized duration for that reading was

$$d-t/norm = (d-t - \mu-s) / \sigma-s \quad (1)$$

The pseudo-normalized duration thus represents the difference between the duration of a word in the text and its mean duration in the control sentences, measured in number of standard deviations. In the rest of this paper, all references to duration refer to this pseudo-normalized duration. (Note that this is not the same as a z-score, because the mean and standard deviation being used are calculated over a sample distinct from the values being normalized.)

For English, the readings of the control sentences were divided into separate groups according to whether the target words were accented or unaccented. The durations and standard deviations were calculated separately for the two groups. The readings of the text sentences were also sorted into two groups, and the mean and standard deviation used in Equation (1) were those for the accented or unaccented productions in the control sentences, depending on whether the reading from the text whose duration was being 'normalized' was accented or unaccented. If a word was accented in one or more readings of the text, but it was accented in less than two readings of the control sentences, then those

productions were excluded from analysis of lengthening, because it was impossible to calculate the pseudo-normalized duration. The same applied for unmatched, unaccented productions. The number of productions that were excluded for this reason were: E1 - 35, E2 - 130, and E3 - 68.

Speech rate. Speech rate was calculated over breath groups, defined as continuous stretches of fluent speech delimited by sentence boundaries and by any intra-sentential pause longer than 150 ms. (This duration was chosen as a minimum for readily identifiable pauses, and has been used as the lower bound of pause duration in other research, such as van Donzel 1999 and Stirling et al. 2001.) All such pauses were silent in the data analyzed here. Note that some pauses between sentences were shorter than 150 ms, but a breath group was always bounded by the beginning or end of the sentence. List numbers and the word “Note”, which occurred as a one-word sentence in the English text, were excluded from analyses of rate.

Rate was expressed as the number of syllables per second over a breath group, calculated as the number of syllables in the breath group divided by its acoustic duration as measured from the waveform. The number of syllables was counted as the number that were audible in each production, as determined by careful listening, rather than the number that might be expected based on phonological structure. For example, speaker E1 frequently elided the middle vowel in the word “document”, producing only two audible syllables. Syllables were counted on this basis because the number of audible syllables seemed to be a more accurate estimate of the amount of sound material produced by the speaker, and thus more appropriate as a basis for acoustic manipulations of rate, than the number of syllables in a word’s phonological structure. As a result, the number of syllables produced in a sentence sometimes varied among the ten recordings of a single speaker.

Because the interest in this study is the nature of the transition from one sentence to the next, the breath groups chosen for analysis were the last in one sentence, and the first in the following sentence. Speakers often produced different numbers of breath groups in different recordings of the same sentence. If a sentence is produced as a single breath group, then the same breath group is analyzed twice, once as the first breath group after a transition, and then as the last breath group before the following transition. The English speakers tended to produce sentences as a single breath group more often than the French speakers. Counting all readings of the sentences, speaker E1 produced 72% of the readings as one breath group, E2: 78%, and E3: 77%, compared to F1: 51% and F3: 62%. (These calculations exclude one-word “sentences” and those in which the breath groups could not be identified, due to speaker error or other problem.) The English speakers also tended to use fewer breath groups in very long sentences than did the French speakers, although the sentences were of similar length in terms of number of syllables.

2.1.5 Statistical analysis

A repeated measures Anova in SAS was used to analyze the data. For both languages, Topic Transition Type (with possible values Topic Shift, Continuation, Elaboration and Text Marker) was an independent factor. For the English data, there were additional independent factors: whether or not the sentence-final word was accented, and whether the final syllable of the sentence-final word was stressed or reduced. The syllable stress factor was included only in the Anovas in which the dependent measure was the pseudo-normalized durations of the target words and of their final rimes.

The dependent measures were duration of the pause following a sentence, speech rate in syllables per second of the last breath group of a sentence and of the first breath group in the following sentence, and the pseudo-normalized durations of the target words and of their final rimes (in English) or syllables (in French). Each dependent measure was analyzed separately for each speaker. The ten recordings of each sentence were treated as the repeated measure variable; measures of different recordings of the same sentence are assumed to be more highly correlated than measures of different sentences. This analysis was run using the Mixed procedure, which is suitable for analyzing data with unequal n’s (Littell et al., 1996). In the Mixed design, the degrees of freedom are different from what might be expected, because the different sources of variance are incorporated directly into the model equation. Where an Anova showed a significant main effect of Topic Transition Type, Scheffé’s test was used for a post-hoc comparison of the means to test which Transition Types differed significantly. Only significant results for Topic Transition

Type are discussed in this paper; complete statistical tables for speakers E1 and E2 are available in Smith (2004).

In addition to the Anovas on the variables listed above, statistical tests were also performed to determine if there was a change in speech rate from the last breath group of a sentence to the first breath group in the following sentence. This was measured by subtracting the speech rate preceding the transition from the speech rate following it, so that a positive number indicates an increase in rate. Transitions where either breath group was a single word were excluded. The difference between the two rates was calculated for each pair of sentences, and a one-sample sign test performed in StatView (SAS Institute 1998) was used to determine whether there was an above-chance number of tokens with an increase or decrease in rate across the transition. Note that this is different from determining whether there is a difference between the mean rates in the breath groups before and after a transition. For the sign tests, statistical significance was tested separately for each transition type.

2.2. Results

2.2.1 English speakers

All three English speakers studied here showed significant effects of Topic Transition Type on all the durational parameters. In general, Topic Shifts were characterized by longer durations than other types of transitions, but the effects varied from speaker to speaker.

Pause duration. All speakers had a significant main effect of Topic Transition Type on pause duration (E1: $F(3,27) = 17.35$, $p < .0001$, E2: $F(3,27) = 95.13$, $p < .0001$, E3: $F(3,27) = 22.84$, $p < .0001$). Post-hoc tests showed that pauses were significantly longer at Topic Shifts for all three speakers. No other differences were significant for speakers E1 and E3. Pauses at Elaborations were significantly shorter than elsewhere for speaker E2.

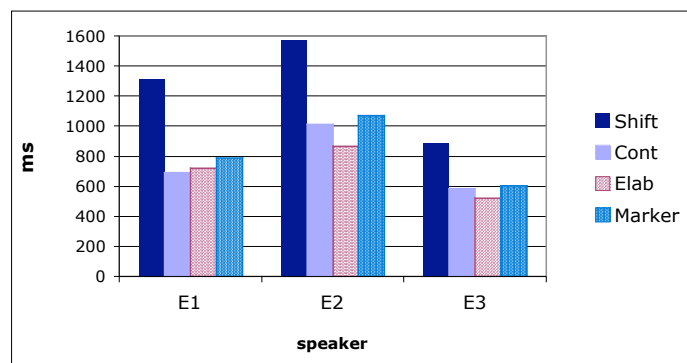


Figure 1. Pause durations for three English speakers.

Final lengthening. The main effect of Topic Transition Type was significant for the amount of lengthening in the sentence-final word for speakers E1 ($F(3,27) = 15.92$, $p < .0001$) and E2 ($F(3,27) = 12.22$, $p < .0001$), but not for speaker E3 ($F(3,27) = 0.25$, ns). However, post-hoc Scheffé tests showed that only a subset of Transition Types differed significantly. There was more lengthening in the final word at Topic Shifts than at Elaborations for both speakers E1 and E2. Speaker E2 also showed significantly more lengthening at Topic Continuations than at Elaborations, but less at Continuations than Shifts. These effects are plotted in Figure 2. All three speakers showed a significant effect of Topic Transition Type on the amount of lengthening in the final rime of the final word (E1: $F(3,27) = 19.24$, $p < .0001$; E2: $F(3,27) = 4.90$, $p < .01$; E3: $F(3,27) = 4.36$, $p < .05$). In the post-hoc tests, Topic Shifts had significantly more lengthening than all other Transition Types for speaker E1; for speaker E2 the difference was significant only between Topic Shifts and Text Markers. Speakers E1 and E3 both had significantly more lengthening at Topic Continuations than at Elaborations or Text Markers. Thus for speakers E1 and E2, both for the final word and for the final rime alone, the pattern was for the most lengthening at Topic Shifts, but speaker E3's Topic Shifts did not differ significantly from other Transition Types.

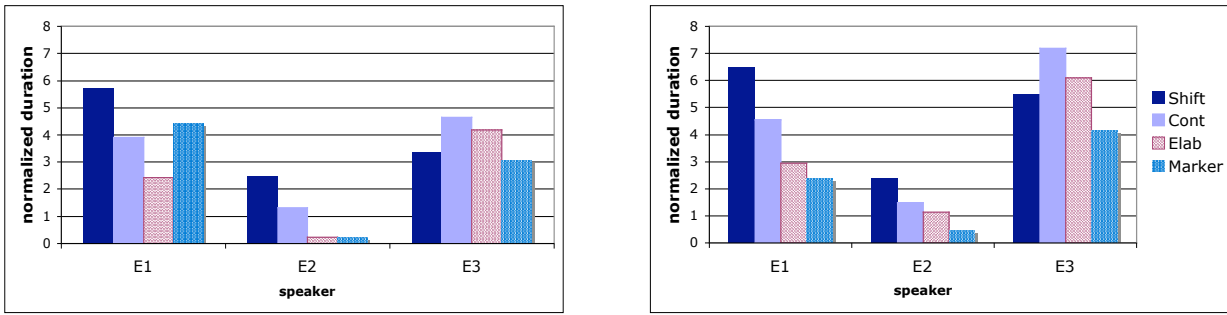


Figure 2. Amount of lengthening, in pseudo-normalized duration, in the sentence-final word (left) and in the final rime of the sentence-final word (right), for three English speakers.

Speech rate in the breath groups surrounding the transition. For each sentence-to-sentence transition, speech rate was measured in the last breath group of one sentence and the first breath group in the following sentence. There were similar patterns for breath groups preceding and following the same Topic Transition Type. All three speakers had significant main effects of Topic Transition Type on speech rate in both breath groups (for the last breath group in the sentence preceding the transition: E1: $F(3,27) = 12.05$, $p < .0001$; E2: $F(3,27) = 16.57$, $p < .0001$; E3: $F(3,27) = 9.01$, $p < .0005$; for the first breath group following the transition: E1: $F(2,18) = 17.31$, $p < .0001$; E2: $F(2,18) = 13.00$, $p < .001$; E3: $F(2,18) = 37.10$, $p < .0001$). Topic Shifts and Continuations had significantly slower rate than Elaborations in both breath groups for all speakers. In the sentence-final breath group, there was also no significant difference between the rates at Elaborations and Text Markers. (In the sentence-initial breath group, rate was not measured in Text Markers, since these were single words.)

The slower rate at Topic Shifts is in accord with previous studies of English which have suggested that utterances beginning new discourse segments (which correspond roughly to utterances following Topic Shifts) tend to have slower speech rates (e.g., Grosz & Hirschberg 1992). However, in the present data, speech rate at Topic Shifts never differed significantly from the rate at Topic Continuations, a result which is at odds with the notion that Topic Shifts represent more major transitions.

Change in speech rate. Whereas the overall speech rate (discussed in the previous paragraph) did not differ between Topic Shifts and Continuations, these Transition Types did differ with respect to rate change. None of the speakers had a significant change in rate at Topic Shifts, but all three showed a significant increase in rate at Topic Continuations (speaker E1: $n=227$, $p < .05$, E2: $n=218$, $p < .01$, E3: $n=180$, $p < .0001$). In addition, speaker E2 decreased rate at Elaborations ($n=145$, $p < .001$).

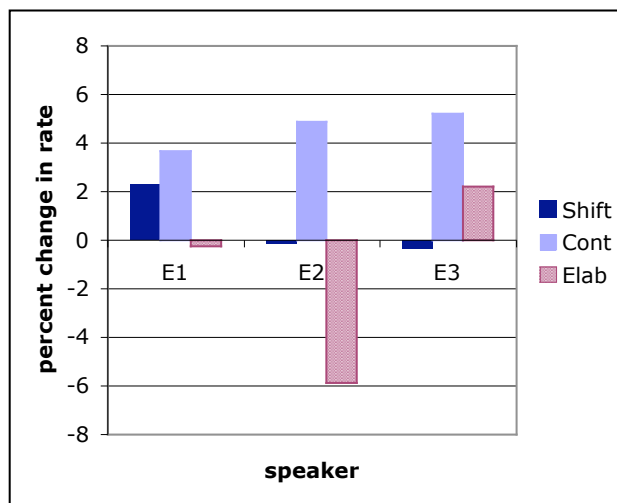


Figure 3: English speakers' percentage change in rate between breath group preceding a transition and breath group following the transition.

2.2.2 French speakers

Both French speakers showed significant effects of Topic Transition Type on some of the durational parameters studied here, but the effects were significant on fewer parameters than was the case for the English speakers, particularly for speaker F2. This may imply that topic organization has a lesser role in durational patterning in French than in English, or it may be that French has less durational variation in general.

Pause duration. Overall, speaker F1 tended to take longer pauses between sentences than speaker F2 (mean for F1: 1095 ms, for F2: 864 ms). Both speakers had a significant effect of Topic Transition Type on pause duration (F1: $F(3,27) = 108.94$, $p < .0001$, F2: $F(3,27) = 81.86$, $p < .0001$). Pauses at sentence boundaries preceding Text Markers were the longest, followed by pauses at Topic Shifts. Post-hoc tests showed that the difference between these was significant for speaker F1 but not for speaker F2. The length of the pauses preceding Text Markers suggests that the speakers were treating these Markers as indicating a new segment of the text. For both speakers, pauses at Topic Continuations were significantly shorter than those at Shifts, and those at Elaborations were shorter still. These results are consistent with the hypothesis that longer pauses are associated with more major discourse boundaries.

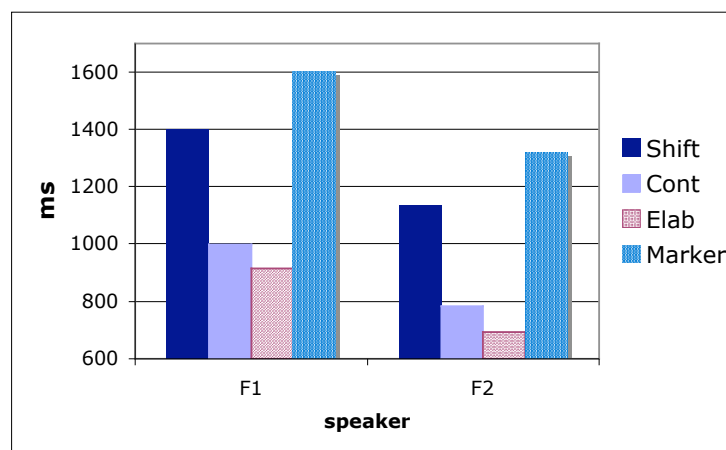


Figure 4: Pause duration for two French speakers.

Final lengthening. The amount of lengthening was measured in the final word of each sentence, and the final syllable of this word. Speaker F1 showed significant effects of Topic Transition Type on both of these measures (word: $F(3,27) = 26.62$, $p < .0001$, syllable: $F(3,27) = 24.61$, $p < .0001$), but speaker F2 did not (word: $F(3,27) = 2.79$, $p = .0595$, syllable: $F(3,27) = 1.48$, $p = .2422$). For speaker F1, there was significantly more lengthening in both word and syllable at Topic Shifts than at other types of transition. The final word's pseudo-normalized duration was 1.94 at Topic Shifts, and the final syllable was 2.35. The least lengthening occurred at Text Markers, with Elaborations and Continuations intermediate and not significantly different. For lengthening in the sentence-final word, there was also no significant difference between Text Markers and Elaborations.

For F1's sentence-final words, the pseudo-normalized duration preceding Text Markers was -0.13 , meaning that these words were in fact slightly shorter in the text than in the control sentences. The final syllable showed just a very slight lengthening preceding Text Markers, with a pseudo-normalized duration of $.16$. This is in contrast to the significant lengthening at Topic Shifts. The difference in amount of lengthening between Text Markers and Topic Shifts contrasts with their similarity in both being followed by long pauses, and suggests that the different durational parameters are patterning independently of one another. An alternative pattern would be a trade-off among the parameters, such that longer words are followed by shorter pauses, and vice versa. Although this is the case for transitions to Text Markers (less lengthening, longer pauses), there does not appear to be a trade-off with other types of transition.

Speech rate. Neither French speaker showed any significant difference among the Topic Transition Types in the speech rate of breath groups preceding or following the transition (speaker F1 preceding the

transition: $F(3,27) = 0.89$, $p = .4576$, following the transition¹: $F(2,18) = 0.08$, $p = .9231$; speaker F2 preceding the transition: $F(3,27) = 2.74$, $p=.0629$, following the transition: $F(2,18) = 2.26$, $p=.1335$). This implies that different types of transition were not associated with significantly faster or slower speech, measured in either of the breath groups.

However, both speakers showed significant differences among Topic Transition Types in the magnitude of the change in speech rate between the breath group preceding the transition and the breath group following it. For speaker F1, the change in rate was significant only at Topic Shifts, where there was a significant slowing in rate ($n=59$, one-sample sign test $p<.01$). Speaker F2 also had significant slowing at Topic Shifts ($n=53$, $p<.01$), as well as an increase in rate at Elaborations ($n=174$, $p<.0001$). The graph in Figure 5 illustrates the average magnitude of the rate change, which is different from the statistical tests of its reliability of occurrence. The figure plots the mean change in rate from the breath group preceding the transition to the breath group following it, calculated as a percentage of the rate in the preceding breath group.

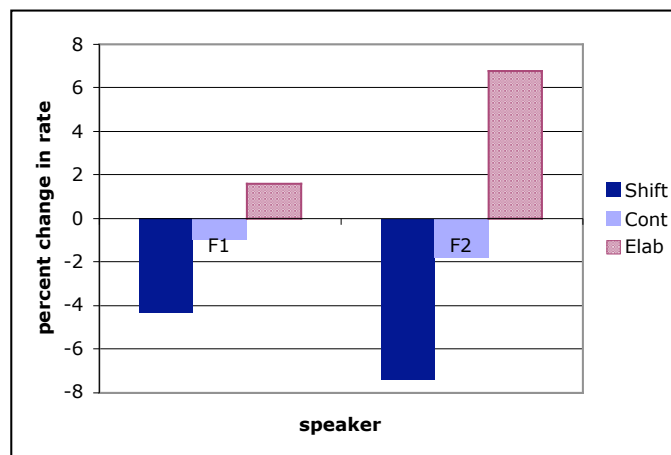


Figure 5: Mean percent change in speech rate from breath group preceding transition to breath group following it, for speakers F1 (left) and F2 (right).

2.2.3 Comparison of the two languages

As is often the case in studies of discourse-level effects in production, the individual speakers varied considerably. The most robust pattern in both languages is that Topic Shifts tend to be characterized by longer pauses, slower speech rate and greater lengthening. There are exceptions to this pattern, such as the lack in French of any significant difference in rate between transition types, but overall this pattern is more reliable than any other in these data.

Results for pauses and lengthening in English and French were fairly similar. In both languages pauses were longer at Topic Shifts than at Continuations or Elaborations. For speakers who had a significant effect of Transition Type on the amount of final lengthening (E1, E2 and F1), lengthening was greater at Topic Shifts than at Continuations or Elaborations (except that the difference between Shifts and Continuations was not significant for speaker E1). These same speakers showed similar effects for lengthening of the final rime (English) or syllable (French), although with significance in fewer comparisons.

There were also some striking differences between the languages, particularly in the measures of speech rate. The English speakers used significantly slower speech rates before and after more major transitions (Shifts, Continuations) than around more minor transitions (Elaborations, Markers). They sped

¹ The degrees of freedom are fewer for the breath group following the transition because there are only three types of transition tested for this variable, rather than four. Text Markers were excluded from this measure because they are single words.

up at Topic Continuations, but maintained an approximately constant rate at Topic Shifts. French speakers, on the other hand, slowed down at Topic Shifts, and did not change rate at Topic Continuations. They also did not have globally different rates at different types of transition. Another difference was in the duration of pauses before Text Markers. In English, these were among the shortest pauses, whereas in French, they were statistically as long (F2) or longer (F1) than the pauses at Topic Shifts, and much longer than pauses at other Transition Types. In both languages, almost all Text Markers were numbers in a list of instructions. The difference in pause length suggests a difference in reading style or interpretation of the structure of numbered lists.

3. Modeling and listener evaluation

3.1. Modeling

The production part of the study demonstrated that there are significant effects of Topic Transition Type on durations in both English and French. Having established that speakers do produce durational variation that may be related to this kind of discourse-level organization, it remains to determine whether this variation is also relevant for listeners. The purpose of the modeling part of this study is to enable testing of the hypothesis that listeners will prefer to listen to speech incorporating the discourse-level durational patterns identified in the production study. Both the systematic effects of Topic Transition Type as well as intra-speaker variability are included in the modeling, since both are hypothesized to contribute to the perception of naturalness.

Using the model developed here, samples of the speech that had been recorded for the production study were modified in several ways to reflect both the systematic patterns and the variability. The different versions of the modified speech were then presented to listeners for evaluation. While the ultimate goal of the research is to propose possible improvements for synthesized speech, modified natural speech was used here in order to preserve the speaker's voice quality and prevent listeners from being distracted by unrelated problems with synthesized speech.

3.1.1. Selection of an appropriate modeling framework

The parameters being modeled are four measures from the production study: duration of the sentence-final word, pseudo-normalized to abstract away from durational properties of individual words, duration of the pause between the two sentences, speech rate in syllables per second of the last breath group of the sentence preceding the transition, and the rate of the first breath group of the sentence following the transition.

The basic approach used for the modeling was to take original recordings and modify them, creating several different versions. The main effect of Topic Transition Type was modeled by taking the mean for each durational parameter for each Topic Transition Type. If speech that is generated using these different mean values is preferred to speech produced without taking Topic Transition Type into account, then it is reasonable to conclude that the effect of Topic Transition Type is relevant to listeners.

It is more challenging to model speaker variability. The goal here is to model the variability of just one speaker, in order to achieve a more comprehensive representation of that speaker's behavior. A possible approach would be to derive a function which approximates the distribution of values for each measured parameter in terms of a small number of parameters that can be readily estimated. This technique has been applied to the modeling of speech durations by, e.g., Crystal and House (1982). The data from the present study do not appear to correspond closely to any well-known distribution, and any such description would inevitably suppress some of the variability that the model intended to capture. Also, the distribution of values varies from speaker to speaker, so a functional description does not seem to offer great advantage in terms of generalizability. Therefore an alternate, non-parametric technique was chosen.

This technique, known as the bootstrap (Efron and Tibshirani, 1993), uses random sampling with replacement to estimate distribution parameters. It makes no assumptions about the structure of the data.

As a statistical technique it is often used for sampling the results of statistical tests. In the context of the present study, bootstrapping makes it possible to model variability without constraining possible values.

3.1.2 Implementation of the model

One recording from speaker E1 and one from speaker F1 were selected as the basis for the modeling. The recordings chosen were those whose mean values for the various measured parameters most closely matched the grand means for that speaker over the entire set of ten recordings. Three shorter passages were extracted from the recording of the text in each language. These passages were selected with two considerations in mind: (i) each passage contained at least one example of the less frequent Transition Types (Marker and Shift) as well as the more common Continuations and Elaborations, with the exception of one French passage which contained no Topic Shifts; and (ii) each passage was more or less self-contained with regard to content. The three passages in English were eight, thirteen and twelve sentences long. The passages in French comprised nine, fourteen and thirteen sentences. The digitized versions of the three passages, as originally recorded by the speaker, will be referred to as the *natural* versions. The other, modified versions were all based on the *natural* version.

Parameter values for the modified versions were calculated using all data for a given speaker that had been analyzed in the production study. The *fixed* version used the grand mean value for each durational parameter, so that any given parameter had the same value in every sentence. This means that in the *fixed* versions of the three English passages, all inter-sentence pause durations were identical, and speech rate was the same in all sentence-final breath groups. In sentences containing more than one breath group, the sentence-initial breath groups all had the same speech rate. The sentence-final words were all lengthened the same amount, proportionate to their sentence-medial durations.

The second of the modified versions, the *mean* version, used the mean value of each durational parameter for each Topic Transition Type; thus, within each language, the same value was used in every sentence of the same Topic Transition Type, but sentences of different Topic Transition Types had different values. This version enables testing of the main effect of Topic Transition Type.

The other modified versions of the passages used bootstrapping techniques to simulate the variability of the speaker's productions. In one version, referred to as the *pause* version, the values for pause duration were bootstrapped (sampled), but for speech rate and the pseudo-normalized word durations, the same values were used as in the *mean* version. The *rate* version bootstrapped values for pseudo-normalized word durations and the two speech rate parameters, with the mean values retained for pause duration. Pseudo-normalized duration and rate were grouped in the modeling because it seemed that manipulating them separately would create anomalous conditions. Finally, the *all* version used bootstrapped values for all four durational parameters.

The sets of data for the two languages were handled separately, but the same methodology was used for both. A full description of the bootstrapping methodology is provided in Smith (2004). Briefly, for the versions which used bootstrapped values, each durational parameter was sampled separately for each Topic Transition Type. One hundred sets of samples were generated (using the RAND function in Microsoft Excel) for each combination of durational parameter and Topic Transition Type, with each set containing the number of samples needed to provide a sampled value for each occurrence of that combination of parameter and Topic Transition Type in the passages being modeled. Out of the one hundred sets of samples, the set whose mean was closest to the grand mean of all the sets was chosen for use in the modeling. This method ensured that (i) all values arrived at through sampling were values that had been measured in the data for that combination of durational parameter and Topic Transition Type; (ii) all measured values were equally likely to be selected through the random sampling, so the sampled values approximated the variability of the measured data; (iii) although no constraints were placed on the variance of the sampled values, their mean closely matched the mean of the subset of data they represented, thus preserving the main effect of Topic Transition Type.

In this way, for the *pause*, *rate* and *all* versions, a set of sampled values was assembled for each of the four durational parameters being modeled. Parameter values were selected for each sentence in each

manipulated version of the passages using an automated procedure in Excel. Then, except for pause durations which were already in appropriate units, arithmetic calculations were performed to transform the pseudo-normalized durations to millisecond values, and convert rate measurements to absolute durations over the set of words that constituted the breath group(s) in each sentence. In the recording that was modeled in English, no sentence contained more than two breath groups, with the result that all words in all sentences were included in the manipulation of rate. In the French recording, out of the 36 sentences that made up the three passages, 13 were produced with three or more breath groups. In these sentences, rate was unchanged in the sentence-medial breath groups.

Once the modeled parameter values were converted to actual durations in milliseconds, these values were applied to the recorded passages. Starting from the digitized, *natural* version of each passage, new waveform files were created in which the durations of the relevant portions of the speech waveform were modified using the automatic Tempo modification command in Macromedia's SoundEdit program. This software was used because, of the several programs evaluated for this purpose, it best preserved voice quality despite changes in rate. For English, five manipulated versions of each passage were created, which resulted in a total of eighteen test passages in all, including the natural version. As will be explained in section 3.2.2 below, the results for English showed that the *all* version was perceived as very poor by listeners. For this reason, no *all* version was created for French, which meant that the total number of French test passages was fifteen.

3.1.3 Correlations

For both the English and French speakers who were modeled, some of the durational parameters were correlated in some Topic Transition Types. It seemed desirable to preserve these correlations in the modeled passages in order to maximize the extent to which the speaker's durational patterns were preserved in the modeling. For simplicity, all correlations that were included were modeled as being perfectly correlated, ignoring differences in the degree of correlation.

Because the distributions of the durational parameters were not normal, the correlations among them were tested using Spearman rank correlations, a non-parametric test which tests for correlation between rank orders, not values (SAS Institute 1998). The number of values used in calculating each correlation varied greatly: from 36 to 246 (mean 134) for English, and from 59 to 394 (mean 166) for French. This number depended on the number of sentences labeled with a given Topic Transition Type, the number of values excluded (e.g., sentences at page breaks were excluded in calculating correlations involving pause duration), and the number of values missing due to factors such as speaker error. Because in most cases the number of values used was fairly large, many of the correlations with $p < .05$ had low correlation coefficients, indicating that little of the variance was explained by the correlation. For this reason, the criterion for significance was made more stringent: $p < .01$ or better, and the correlation coefficient $\rho > .2$. Correlations that met these criteria are listed in Table 2.

The correlations listed in the table were needed for the modeling of the *rate* and (in English) the *all* versions of the passages. For French, of the correlations listed as significant in Table 2, the only ones needed for modeling were those involving the duration of the final word and speech rate, since the decision not to create an *all* version in French meant that correlations between pause duration and other parameters were not modeled. Unfortunately, due to experimenter error, the correlations were not implemented correctly in the modeling of French. The correlation at Topic Shifts between duration of the final word and speech rate in the first breath group of the 2nd sentence was incorrectly omitted, and four other correlations were erroneously included. These resulted in measures of speech rate in Continuations, Elaborations and Text Markers that were incorrectly modeled as being correlated. Because the samples should have been randomly paired rather than correlated, it is impossible to determine a specific value for the extent of the error involved. However, comparison of the pairings used for each Topic Transition Type with possible random pairings shows that the average absolute value of the differences in speech rate ranged from 0.2 to 1.4 syllables/second. Since the error was confined to the correlations among the values, and did not affect the selection of values used, it seems unlikely that it had any major consequences for the perceptible amount of variation in rate.

Parameter pairs	Speaker E1	Speaker F1
pseudo-norm. duration final word – pause duration		Topic Shifts Continuations Elaborations
pseudo-norm. duration final word – speech rate last breath group 1 st sentence	Elaborations Text Markers	Continuations
pseudo-norm. duration final word – speech rate first breath group 2 nd sentence		Topic Shifts Elaborations
p a u s e d u r a t i o n – speech rate last breath group 1 st sentence		Elaborations
p a u s e d u r a t i o n – speech rate first breath group 2 nd sentence		Topic Shifts
speech rate last breath group 1 st sentence – speech rate first breath group 2 nd sentence	Continuations Elaborations	

Table 2: Topic Transition Types in which there were Spearman rank correlations significant with $\rho > .2$ and $p < .01$ or better.

3.2. Listening tests

3.2.1 Method for the listening tests

The test passages produced from the modeling were presented to listeners for ranking on a seven-point scale, using the program PsyScope (Cohen et al., 1993) for control of the experiment, which was conducted using a Macintosh PowerBook computer. Each listener was tested individually. For the French listeners, all experimental materials and interaction with the experimenter were in French. The experimental instructions and task were presented as a series of screens; listeners controlled the pace at which they moved through the experiment. They were first presented with instructions, and then heard recordings of two example passages. These passages were different from any of the test passages but had been extracted from the same original recording. The example passages were each followed by a screen explaining that one of the example passages was very poor and one very good. These examples were intended to familiarize listeners with the extremes in quality of the passages they would hear. The instruction phase concluded with additional passages which the listeners were asked to rate for practice with the procedure. These were followed by a screen explaining that these practice passages were intermediate in quality and should therefore receive ratings in between the two extremes. The purpose of providing a suggested range for ratings of the practice passages was to familiarize listeners with the variability of the manipulations and encourage them to use the full range of possible ratings.

For the English experiment, fifteen listeners participated. They were recruited by flyers posted around the campus of the University of New Mexico, where they were tested. Twelve French listeners were recruited and tested at the University of California, Berkeley. All listeners were paid \$10 for their participation. The tests took approximately half an hour to complete.

Since preliminary analysis showed no significant differences among the three passages of text in each language, the listeners' ratings were averaged over the three passages, then rank-ordered, giving a total of six scores per English listener and five for each French listener (one for each version). These scores (rank orders) ranged from 1 (lowest mean rating) to 6 (highest mean rating) for English, and from 1 to 5 for French. Statistical analyses were computed on these rank-order scores. The degree of agreement among the listeners was assessed by computing the kappa statistic (Fleiss, 1971). For English this had a

value of .56, within the range described by Landis and Koch (1977) as indicating “moderate” agreement. For the French listeners, kappa was equal to .29, indicating ‘fair’ agreement.

3.2.2 Results of the listening tests

English. The six versions differed significantly ($\chi^2 = 66.6$, $p < .001$) when analyzed in a Friedman test (a non-parametric two-way Anova) in StatView. Every listener rated the *natural* (unmodified) version as the best, and 14 of the 15 listeners rated the *all* version as the worst. The graph on the left of Figure 6 shows the scores averaged over the entire group of listeners. Pairwise comparisons were performed among the six versions’ scores using the Bonferroni test in the SAS program GLM. All of the pairwise comparisons were significant at the .01 level except for the *mean* and *pause* versions, which did not differ significantly.

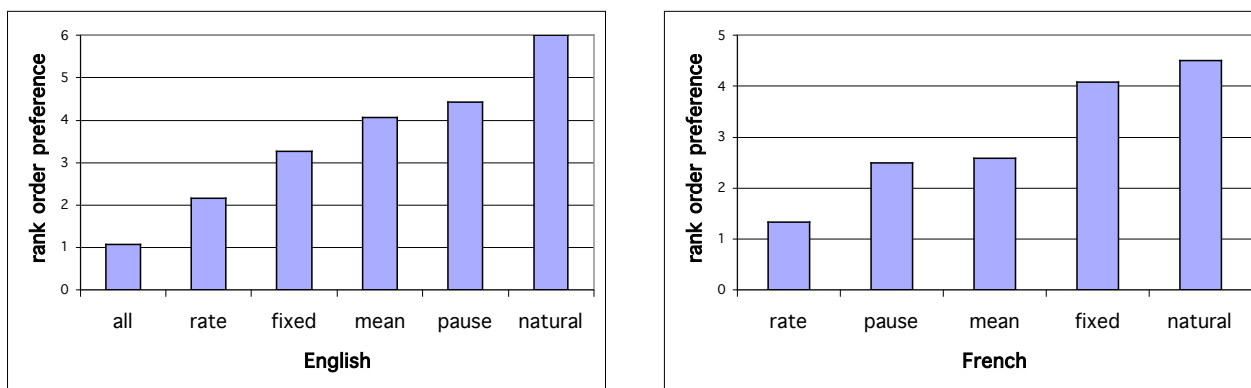


Figure 6: Rank orders of listeners’ preferences for the different versions of English (left) and French (right), averaged over three test passages.

French. As was the case for English, the Friedman test showed that the five French versions differed significantly ($\chi^2 = 33.5$, $p < .001$). In other regards the results for French (shown on the right of Figure 6) differed from those obtained for English. Pairwise comparisons using the Bonferroni test in SAS GLM showed that there was no significant difference between listeners’ rankings of the *natural* and *fixed* versions. These two were rated significantly better ($p < .01$) than the *mean* and *pause* versions, which were statistically equivalent. Seven of the twelve listeners rated the *natural* version as best, with two others ranking it in a tie for best. The *rate* version was rated lowest overall, and by ten of the twelve individual listeners.

3.2.3 Comparing the results in the two languages

The English and French listeners showed quite different preferences. The English listeners preferred the *mean* and *pause* versions, in which the durations were modified to include differences among the Topic Transition Types, over the *fixed* version, which did not differentiate among Transition Types. This is compatible with the original proposal that incorporating topic-related variation into synthetic speech would improve its quality (with the proviso that the speech tested here was not synthetic). In contrast, the French listeners preferred the *fixed* version to the *mean* and *pause* versions. In other words, they preferred an absence of variation in the durational parameters to the presence of variation. This suggests that their preference in synthetic speech would be for a more constant rhythm, not for including variation reflecting topic structure.

In other respects, the listeners of the two language groups responded more similarly. For both languages, there was no statistical difference between listeners’ rankings of the *pause* version, with random sampling of pause duration, and the *mean* version. Recall that these two versions shared the same parameter values for word duration and rate. Changes in pause durations alter only the duration of the silent interval, which is much less salient acoustically than the manipulations used in the *all* and *rate* versions. Listeners in both languages disliked the *rate* version, and in English, the *all* version, both of which involved random sampling of speech rate and the duration of the sentence-final word. The varying rate and word duration in these two versions would have affected the rate of change in the formant trajectories, potentially altering the sound qualities quite substantially and not in accordance with the

spectral changes normally associated with prosodic effects (Wouters and Macon 2002). It is thus perhaps not surprising that these two versions were disliked. The similarly low ranking given by English and French listeners to these versions suggests that it is very unlikely that the errors in modeling the correlations in the *rate* version for French, as noted in section 3.1.3, are responsible for the low ranking of this version.

It is possible that topic-related variation produced in a different way might be pleasing to French listeners. The French listeners' preference for versions without durational modifications may be explained, at least in part, by the modeling procedure being less successful in capturing important aspects of durational variation in French than in English. One significant pattern observed in French was the slowing of rate from one sentence to the next at Topic Shifts, without there being a correlation between the two rates. This was not captured in the modeling because rate was specified separately for the breath groups preceding and following each transition. The modeling of rate interacted with another difference between the languages. Of the 33 sentences included in the passages for the English listening task, 30 consisted of a single breath group, so rate did not vary within the sentence. In contrast, most of the French sentences (20 out of 36) had two or more breath groups, whose rates was manipulated independently in the modeling. It is likely that an abrupt change of rate in the middle of a sentence is disruptive to a listener. In natural speech, rate may change gradually, but there was no mechanism in the modeling for achieving this. In this way, modifications to rate may have been more problematic in French, because the recording being modeled in French had more sentences that were made up of multiple breath groups. This may have led French listeners to give lower ratings to versions involving rate changes, favoring the *fixed* version which kept a constant rate throughout.

A different reason that the modeling may have been less successful for French is that it relied crucially on the labeling of topic structure, on which there was less agreement among the French labelers than among the English labelers. If the lower rate of agreement also holds for other French readers of the text, then it is less likely for French than for English that the speakers' and listeners' interpretation of the text will match the labeling. If the speakers interpret the transitions differently than what is implied by the labeling, then the durational modifications are less reliably related to the organization of the text. If the listeners interpret the text differently, they may be less satisfied by the durational modifications that they hear.

4. Discussion

In productions in both languages, the measured durational parameters differed significantly among different Topic Transition Types, suggesting that topic organization is a factor that contributes to determining speaker durations in both English and French. But these languages clearly differ in the way in which topic organization, as analyzed here, is produced and perceived. The measures of speech rate, in particular, showed interesting differences. In English, all speakers spoke significantly more slowly before and after Topic Shifts and Continuations, but there was no consistent pattern of *change* in rate between the breath groups preceding and following the transitions. In contrast, both French speakers had a significant pattern of slowing down at Topic Shifts, but neither speaker showed any effect of Topic Transition Type on mean rate in either of the breath groups preceding or following the transitions.

It is possible that this apparent difference between the languages is at least in part a consequence of a difference in the number of breath groups per sentence. The English speakers tended to produce a sentence as one breath group. This means that the rate produced in that breath group was treated as following one transition and preceding another. If these transitions were labeled with different Transition Types, then the rate measured in that breath group is averaged into the mean for two different Transition Types. The French speakers were more likely to use multiple breath groups in a sentence, although even for them, most sentences contained only one breath group. In sentences with multiple breath groups, the rate in a breath group is associated with just one sentence-to-sentence transition. This means that rate

changes at one type of transition will not be “diluted” by the influence of other transitions. It may therefore be more likely that small changes between breath groups can reach significance.

Based on the results of the listening tests, the method of introducing variability that was tried here was not a success. Listeners in both languages disliked the versions with random sampling of lengthening or rate, even though the values used were all ones that the speaker had produced for the appropriate Transition Type. Rather than concluding that listeners do not like variability, a more judicious conclusion may be that listeners like variability that is appropriate for the context. This involves many factors, of which this study has considered only one. Modeling variability as largely random is inappropriate for the great part of it which is not random.

It is not entirely clear whether the differences between the languages that were observed in this study are best described as localized or as global, but on balance they seem to be more global. A localized durational effect is one which is evidenced in a narrow time window; a global effect is one that is distributed over multiple linguistic units. The largest difference between the languages observed in the present study was in the listening results. These reflect listeners’ holistic responses to the different versions and therefore seem likely to reflect global differences. In the production study, the differences observed in speech rate also seem more global. On the other hand, pause duration, the most localized parameter measured here, patterned the most similarly in English and French.

The results of this study seem to concur with Vaissière’s (1983) description of timing in English and French as using the same elements but with different priorities. She describes English as having a stress-based rhythm, and French a temporal rhythm based on lengthening before boundaries² (Vaissière 1997). This is undoubtedly true at the word and accent group level. But in this study there has been little evidence in French of final lengthening at boundaries between larger, discourse-level units. In terms of the Transition Types used in this study, Topic Shifts correspond to a boundary between larger units than do the other Transition Types. The pause duration results suggest that the French speakers may also have treated Text Markers as signaling a major boundary, since there were long pauses at Text Markers. But while the words preceding Topic Shifts showed considerable lengthening for one French speaker (as well as the English speakers) there was no lengthening preceding Text Markers in French. The pause durations and lengthening appear to vary independently, as suggested by Lehiste (1982). Also, the rate results for French showed slowing following a Topic Shift boundary, which is at odds with the notion of lengthening before the boundary. These results make it difficult to interpret the French discourse-level durational patterns seen in this study as concordant with a system that signals boundaries by pre-boundary lengthening. But it is also possible, given that the French listeners did not like speech modeled with the transition-related durational patterns observed here, that in some way the data measured and modeled here did not reflect the rhythmic expectations of French language users.

The results for English in this study are in general agreement with results of previous studies such as those cited in the introduction, in that Topic Shifts, marking larger boundaries, were generally associated with longer durations. Most importantly, English listeners preferred to hear speech which instantiated this pattern, suggesting that it may be part of their unconscious knowledge or set of expectations of the language. Thus the original hypothesis of this study, that the introduction of these patterns offers a potential way to improve synthesized speech, seems justified for English. However, as noted above, this success is confined to the pattern of main effects, not the pattern of variability which was undoubtedly over-simplified and thus not appropriate for the different contexts that occurred in the modeled speech.

Clearly much remains to be done in elucidating the differences between these two languages at the discourse level, even though this study treated only one, very constrained style of speech. It also seems essential to continue to seek connections between prosody at the word and phrase level with prosody at higher levels, even though from this study the different levels seem to operate rather differently, at least for French. Exactly how the different units of speech that contribute to timing are related, and to what extent patterns of discourse-level timing are language-specific, remains to be determined.

² « basé sur l’allongement en fin des unités » (Vaissière 1997:54)

Acknowledgements

This work was supported by NSF grant BCS-9983106. It would not have been possible without the contributions of the speakers, listeners and labelers, to whom I am very grateful. And many thanks to Lisa A. Hogan and Monica Valandani for research assistance.

References

- ASTÉSANO C. (2001), *Rythme et Accentuation en français : Invariance et Variabilité Stylistique*, Paris, L'Harmattan
- AYERS G. (1994), Discourse functions of pitch range in spontaneous and read speech, *OSU Working Papers in Linguistics 44*, 1-49
- BERTINETTO P-M (1989), Reflections on the dichotomy «stress» vs. «syllable-timing», *Revue de Phonétique Appliquée*, 91/93, 99-130
- COHEN J., MACWHINNEY B., FLATT M. & PROVOST J. (1993), PsyScope: A new graphic interactive environment for designing psychology experiments, *Behavioral Research Methods, Instruments, and Computers 25*, 257-271
- CRYSTAL T. & HOUSE A. (1982), Segmental durations in connected speech signals: Preliminary results, *Journal of the Acoustical Society of America 72*, 705-716
- DAUER R. (1983), Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics 11*, 51-62
- DELAIS E. (1994), Rythme et structure prosodique en français, in Lyche C. (ed.), *French Generative Phonology: Retrospective and Perspectives*, Salford, UK, Association for French Language Studies, 131-150
- DI CRISTO A. (1998), Intonation in French, in Hirst D. & Di Cristo A. (eds.), *Intonation Systems: A Survey of Twenty Languages*, Cambridge, Cambridge University Press, 195-218
- DI CRISTO A. & HIRST D. (1997), L'accentuation non-emphatique en français : stratégies et paramètres, in PERROT J. (ed.), *Polyphonie pour Iván Fónagy*, Paris, L'Harmattan, 71-101
- DUEZ D (1999), Effects of articulation rate on duration in read French speech, in Olay G., Németh G. & Erdőhegyi K. (eds.), *Proceedings of Eurospeech 99*, Budapest, ESCA, vol. 2, 715-718
- EDWARDS J., BECKMAN M. & FLETCHER J. (1991), The articulatory kinematics of final lengthening, *Journal of the Acoustical Society of America 89*, 369-382
- EFRON B., TIBSHIRANI R. (1993), *An Introduction to the Bootstrap*, New York, Chapman & Hall, 45-53
- FANT G., KRUCKENBERG A. & NORD L. (1991), Durational correlates of stress in Swedish, French and English, *Journal of Phonetics 19*, 351-365
- FLEISS J. (1971), Measuring nominal scale agreement among many raters, *Psychological Bulletin 76*, 378-382
- FON Y-J. J. (2002), *A Cross-Linguistic Study in Syntactic and Discourse Boundary Cues in Spontaneous Speech*, Ph.D. dissertation, Columbus, Ohio, The Ohio State University
- FON J. (2004), Perception of discourse boundaries by Taiwan Mandarin speakers, in Bel B. & Marlien I. (eds.), *Proceedings of Speech Prosody 2004*, Nara, Japan, 709-712

- GELUYKENS R. & SWERTS M. (1994), Prosodic cues to discourse boundaries in experimental dialogues, *Speech Communication* 15, 69-77
- GROSJEAN F. & DESCHAMPS A. (1975), Analyse contrastive des variables temporelles de l'anglais et du français : vitesse de parole et variables composantes, phénomènes d'hésitation, *Phonetica* 31, 144-184
- GROSZ B. & HIRSCHBERG J. (1992), Some intonational characteristics of discourse structure, *Proceedings of the 2nd International Conference on Spoken Language Processing*, Banff, Canada, 429-432
- HERMAN R. (2000), Phonetic markers of global discourse structures in English. *Journal of Phonetics* 28, 466-493
- HIRSCHBERG J. & NAKATANI C. (1996), A prosodic analysis of discourse segments in direction-giving monologues, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 286-293
- HIRST D. & DI CRISTO A. (1984), French intonation: a parametric approach, *Die Neueren Sprachen* 83, 554-569
- JUN S.-A & FOUGERON C. (2000), A phonological model of French intonation, in Botinis A. (ed.), *Intonation: Analysis, Modeling and Technology*, Dordrecht, Kluwer, 209-242
- KREIMAN J. (1982), Perception of sentence and paragraph boundaries in natural conversation, *Journal of Phonetics* 10, 163-175
- LACHERET-DUJOUR A. & BEAUGENDRE F. (1999), *La prosodie du français*, Paris, CNRS Editions
- LANDIS J. & KOCH G. (1977), The measurement of observer agreement for categorical data, *Biometrics* 33, 159-174
- LEHISTE I. (1979), Perception of sentence and paragraph boundaries, in Lindblom B. & Öhman S. (eds.), *Frontiers of Speech Communication Research*, London, Academic Press, 191-201
- LEHISTE I. (1982), Some phonetic characteristics of discourse, *Studia Linguistica* 36-2, 117-130
- LITTELL R., MILLIKEN G., STROUP W. & WOLFINGER R. (1996), *SAS System for Mixed Models*, Cary, NC, SAS Institute
- NAKAJIMA S., ALLEN J. (1993), A study on prosody and discourse structure in cooperative dialogues, *Phonetica* 50, 197-210
- NOORDMAN L., DASSEN I., SWERTS M. & TERKEN J. (1999), Prosodic markers of text structure in van Hoek K., Kibrik A. & Noordman L. (eds.), *Discourse Studies in Cognitive Linguistics: Selected Papers from the 5th International Cognitive Linguistics Conference*, Amsterdam, John Benjamins, 133-148
- PASDELOUP V. (1992), A prosodic model for French text-to-speech synthesis: a psycholinguistic approach, in BAILLY G., BENOIT C. & SAWALLIS T. (eds.), *Talking Machines: Theories, Models, and Designs*, Amsterdam, Elsevier, 335-348
- POST B. (2000), *Tonal and Phrasal Structures in French Intonation*, The Hague, Thesus
- RAMUS F., NESPOR M. & MEHLER J. (1999), Correlates of linguistic rhythm in the speech signal, *Cognition* 73, 265-292
- SAS Institute (1998), *StatView Reference Manual*, Cary, NC, SAS Institute

- STIRLING L., FLETCHER J., MUSHIN I. & WALES R. (2001), Representational issues in annotation: using the Australian map task corpus to relate prosody and discourse structure, *Speech Communication* 33, 113-134
- SWERTS M. (1997), Prosodic features at discourse boundaries of different strength, *Journal of the Acoustical Society of America* 101, 514-521
- SWERTS M. & GELUYKENS R. (1994), Prosody as a marker of information flow in spoken discourse, *Language and Speech* 37, 21-43
- TURK A. & WHITE L. (1999), Structural influences on accentual lengthening in English, *Journal of Phonetics* 27, 171-206
- VAISSIÈRE J. (1983), Language-independent prosodic features, in Cutler A. & Ladd D.R. (eds.), *Prosody: Models and Measurements*, Berlin, Springer, 53-66
- VAISSIÈRE J. (1991), Rhythm, accentuation and final lengthening in French, in Sundberg J., Nord L. & Carlson R. (eds.), *Music, Language, Speech and Brain: Proceedings of an International Symposium at the Wenner-Gren Center, Stockholm, Basingstoke, UK, Macmillan Press*, 108-120
- VAISSIÈRE, J. (1997), Langues, prosodie et syntaxe, *Revue Traitement Automatique des Langues* 38-1, 53-82
- VAN DONZEL M. (1999), *Prosodic Aspects of Information Structure in Discourse*, The Hague, Thesus
- WICHMANN A. (2000), *Intonation in Text and Discourse: Beginnings, Middles and Ends*, Harlow, Pearson Education
- WIGHTMAN C., SHATTUCK-HUFNAGEL S., OSTENDORF M. & PRICE P. (1992), Segmental durations in the vicinity of prosodic phrase boundaries, *Journal of the Acoustical Society of America* 92, 1707-1717
- WOUTERS J. & MACON M. (2002), Effects of prosodic factors on spectral dynamics. I. Analysis, *Journal of the Acoustical Society of America* 111, 417-427