Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Speakers advance-project turn completion by slowing down: A multifactorial corpus analysis

Christoph Rühlemann [a,*], Stefan Th. Gries [b]

[a] *Albert-Ludwigs-University Freiburg, German and Philipps-University Marburg, Germany*
[b] *University of California, Santa Barbara, United States and Justus Liebig University, Giessen, Germany*

ARTICLE INFO

Article history:
Received 16 April 2019
Received in revised form 21 February 2020
Accepted 4 March 2020

Keywords:
Duration
Turn transition
Turn-final lengthening
Rallentando
British National Corpus

ABSTRACT

Turn transition in talk-in-interaction is achieved with remarkable precision, most commonly following a gap of no more than 200 ms (e.g., Stivers et al., 2009). How the precision is achieved is a complex issue given the wide range of variables co-participants to talk-in-interaction deploy to project (as speakers) and predict (as listeners) turn completion. This paper aims to contribute to a deeper understanding of one such variable used by speakers to project turn-completion: changes in word duration in turns-at-talk. As word duration varies significantly due to influences from a large number of confounds, we approach the challenges inherent in "[p]roviding robust, quantified, comparative measures of duration" (Local & Walker, 2012: 259) by fitting mixed-effects models based on naturally occurring corpus data. Contrary to previous research, which hailed the turn-final drawl as a turn-yielding cue, the models indicate that drawling, or rallentando, affects not just the turn-final syllable/word but large portions of the turn. Rallentando appears to be, not a one-off cue marking the turn's end-point upon its occurrence, but an extended process advance-projecting the turn's durational envelope. Also, as a graded advance-projecting resource, rallentando is in and of itself insufficient to signal turn completion reliably; listeners are likely to rely on turn rallentando *in unison* with other, preferably discrete cues marking the turn-completion point upon its occurrence, for "recogniz[ing] that a turn is definitely coming to an end" (Levinson & Torreira, 2015: 12) and triggering the launch of the next turn.

1. Introduction

A fundamental fact of every-day talk-in-interaction is that "[t]ransitions (from one turn to a next) with no gap and no overlap between them are common'' (Sacks, Schegloff, & Jefferson, 1974: 700). Sacks et al.'s reference to 'no gap and no overlap', however, is not to be taken literally, as an exact acoustic quantification. As Heldner and Edlund (2010: 564) have shown, cases of actual zero gap and zero overlap represent only a ''marginal part'' of turn transitions. In fact, transition commonly happens within a 'slight gap' (Heldner & Edlund, 2010: 557; cf. also Sacks et al. [1974: 700–701]) approximating the average length of a single syllable (Levinson, 2016: 7), which is 200 ms—that is, turn transition happens "faster than the average time it takes to blink the eye" (Enfield, 2017: 41). The 200 ms transition time has subsequently been confirmed in large-scale empirical work. Investigating turn transition times in ten unrelated languages, including, for example, Danish, Lao, Japanese and Tzeltal, Stivers et al. (2009: 10588) discovered "a unimodal distribution with a mode offset for each language between 0 and 200 ms". Working on natural conversations and task-guided talk in three European languages (Dutch, Scottish English, and Swedish) Heldner and Edlund (2010) found that "[t]he most common between-speaker interval (...) is a gap of about 200 ms" (Heldner & Edlund, 2010: 564). Thus, the turn-taking system in human face-to-face interaction seems "strongly universal, with only slight variations in timing" (Levinson, 2016: 7): across languages and cultures, conversationalists time the transition from one turn to the next with remarkable precision.

An intriguing question is *how* this precision is brought about. The mechanisms underlying this precision are complex as conversationalists exploit a wide range of resources to predict (as listeners) and project (as speakers) turn completion.

Drawing on these resources as well as the turn's emerging semantic, pragmatic and perceptual signals listeners form a 'multimodal gestalt' (Holler & Levinson, 2019: 6) to "identify or predict the speech act or action being carried out – both

* Corrresponding author at: Albert-Ludwigs-Universität Freiburg, Deutsches Seminar – Germanistische Linguistik, Platz der Universität 3, 79085 Freiburg, Belfortstr. 18, Germany.
E-mail address: christoph.ruehlemann@germanistik.uni-freiburg.de (C. Rühlemann).

https://doi.org/10.1016/j.wocn.2020.100976
0095-4470/© 2020 Elsevier Ltd. All rights reserved.

the illocutionary force and the likely propositional content" (Levinson & Torreira, 2015: 13).[1] Recent EEG experiments corroborate the assumption that listeners start to predict speech acts *early*, that is, far ahead of the turn's actual completion: Magyari et al. (2014) observed a beta desynchronization 1200 ms before the (lexico-grammatically predictable) end of a turn; in Gisladottir et al. (2018), the processing of three distinct speech acts —answer, declination, and offer—performed by the same utterance ("I have a credit card") started to differ within the first 400 ms of the turn-in-progress.[2]

Speakers, on the other hand, both advance-project turn completion as well as mark completion on its occurrence (cf. Clayman, 2013: 151). Advance-projection relies on morphosyntax and intonation contour, with the former "provid[ing] most of the early clues to the overall structural envelope (e.g., turns beginning with *if* or *either* or *whenever* project a two clause structure), so offering some long distance projection" (Levinson & Torreira, 2015: 13; cf. also Sacks et al., 1974; Clayman, 2013: 158; Magyari et al. 2014: 2538). Once speakers arrive at the turn's completion point they have at their disposal a large number of turn-final cues to give the ultimate 'go-signal' (Barthel, Meyer, & Levinson, 2017: 9) to the recipient to launch their response. The inventory of turn-final cues is extensively multimodal. Turn-final cues may be post-completers, such as address forms or question tags, in which case they are linguistic. Perhaps more commonly, however, turn-final cues are of a non-verbal nature, including gaze, body motion, pauses as well as a large range of vocal cues ranging from audible creaky voice quality, outbreaths to intensity, pitch, and duration, specifically the lengthening of the final word/syllable. An attempt at an exhaustive list of turn-final cues reported in the literature is this:

linguistic:

- syntactic completion: Duncan (1972); Sacks et al. (1974); Wells and MacFarlane (1998); de Ruiter, Mitterer, and Enfield (2006), Levinson and Torreira (2015)
- question tags: Sacks et al. (1974)
- address terms: Jefferson (1973), Sacks et al. (1974)
- idioms: Duncan (1972) ('sociocentric sequences': "stereotyped expressions, typically following a substantive statement" (p. 287) such as 'you know', 'or something', etc.)
- lexico-syntactic predictability: Magyari et al. (2014)
- pragmatic: Wells and MacFarlane (1998), Levinson and Torreira (2015)

nonlinguistic:

- body motion: Duncan (1972) (esp. termination of hand movement)
- gaze: Kendon (1967), Bavelas, Coates, and Johnson (2002) (re-establishment of 'mutual gaze')

paralinguistic:

- creak: Ogden (2001) (for Finnish)
- aspiration of word-final plosives: Local and Walker (2012)
- audible outbreath: Local and Walker (2012), Torreira, Bögels, and Levinson (2015)
- final major accented syllable: Wells and McFarlane (1998)
- pitch drop: Beattie, Cutler, and Pearson (1982); Duncan (1972, 1974); Bögels and Torreira (2015)
- intensity drop ('diminuendo'): Duncan (1972, 1974); Duncan and Niederehe (1974); Gravano and Hirschberg (2011)
- turn-final lengthening ('drawl'): Duncan (1972, 1974); Duncan and Niederehe (1974); Local and Walker (2012); Bögels and Torreira (2015)

The turn-final cue we are centrally concerned with in this paper is turn-final lengthening, also referred to as 'drawl.' While the drawl is firmly established as a primary correlate of constituent structure, with lengthening occurring near constituent boundaries (Turk & Shattuck-Hufnagel, 2007), research into lengthening as a turn-yielding cue has been fraught with difficulties and limitations. One limiting factor relates to the type of data used, which were either highly specialized, experimental, or 'small' in size. For example, Duncan's (1972) analysis is based on two 19-minute excerpts from psychotherapeutic interviews; Bögels and Torreira (2015) conducted a button-press experiment with a small set of paired utterances (e.g., "So you're a student?" and "So you're student at Radboud University?") presented in four manipulated conditions; and Local and Walker (2012) analyzed a naturally occurring, 12-minute short telephone call between two adult speakers of British English. In other words, turn-final drawling has so far not been investigated in naturally-occurring data of scale.[3]

The reason why is obvious: "[p]roviding robust, quantified, comparative measures of duration is problematic when working with naturally occurring materials: syllable and word structure, accentual patterning, position in utterance, speaker, overall speaking rate, information structure etc., are all things which cannot be controlled for and which, moreover, are known to impact on the durational characteristics of words and parts of words" (Local & Walker, 2012: 259).

Another limitation to previous research is of a methodological nature. The established method used to measure turn-final drawls has been to compare the durations of one and the same word in two conditions: (i) when the word occurs inside a turn, that is, *not* as the turn-final word, and (ii) when it is used as the turn-final word; cf., for example, Bögels and Torreira's (2015) experiment with the two utterances "So you're a student?" and "So you're student at Radboud University?", where the word *student* is used turn-finally and, respectively, turn-medially. Deploying this method, Local and Walker (2012: 260) observe that "final tokens [of the same word] are on average 65% longer than medial tokens."

This method begs questions. First, in any corpus of natural conversation the number of word types and word tokens

[1] See also Bar's notion of 'recognition-by-analogy' (Bar, 2009), a three-stage process from (i) sensory input (i.e., the multimodal turn-in-progress) to (ii) analogy based on associations with past experiences and memories (i.e., the speech act performed in turns similar to the turn-in-progress) to (iii) prediction (i.e., the speech act ascribed via analogy to the turn-in-progress).

[2] As an anonymous reviewer cautioned, none of these studies suggest that listeners predict the exact speech act early. What the studies show is that the brain starts working toward deciphering the unfolding speech act far ahead of the speech act's completion.

[3] The reference to the limited set of linguistic stimuli used in Bögels and Torreira (2015) is not to be understood as a critique. Using small numbers of stimuli is inevitable in experimental settings and, in the case of Bögels and Torreira (2015), a purposeful design feature related to the concern, as an anonymous reviewer remarked, that "presenting too many manipulated stimuli could lead participants to react unnaturally or develop strategies specific to the experiment."

occurring in either condition is seriously limited to high-frequency word types, that is, to function words, inserts, and highly frequent lexical words; most lexical words, by contrast, which tend to have very few tokens in any corpus, will be unlikely to occur in both conditions *en masse*.[4] Second, it is unclear as to how listeners should be able to recognize a drawl as a turn-final drawl and, hence, a turn-yielding cue. The lengthening of a syllable or word can only be recognized *contrastively*, that is, as a duration that is longer than some other duration. The question is, precisely longer than *what* other duration? Two different approaches are conceivable. One is what might be called the *lexicon hypothesis*, the other the *speech-rate hypothesis*.

The lexicon hypothesis would hold that speakers/listeners have stored default or canonical durations for words. This hypothesis would then stipulate that listeners recognize that words they are hearing from a speaker are longer than the default durations they have stored for those words and that, within some overall context, this might be a signal to them that a turn is about to end or ending. This hypothesis might seem particularly attractive to scholars coming from a usage/exemplar-based perspective, who assume that speakers store large amounts of usage events in a multidimensional kind of exemplar space and that speakers constantly make (implicit) comparisons between their current input and their past input (as when speakers decide which vowel a certain input sound represents, given its formants in vowel space).

While this hypothesis seems attractive, it will require modification if only because we know that listeners can adjust quickly to (speaker-)specific characteristics of their current input. For instance, we know from research on phonetic accommodation that listeners can adjust their expectations and 'comprehension computations' when they talk to speakers of a different dialect or non-native speakers (see, e.g., Kraljic, Brennan, & Samuel, 2008, Kim, Horton, & Bradlow, 2011). In a sense, they adjust the input to fit the multidimensional space in which they are processing the input so that, after even only small amounts of input, their comprehension works fairly effortlessly again. Thus, a more realistic approach has listeners rely less on comparisons of the current input to their long-term storage of the same words, but also, and more so, to characteristics of the current input. Taking this approach, we propose the speech-rate hypothesis. It holds that speakers decrease the speed of speaking over the course of the turn to advance-signal turn-completion. Listeners, on the other hand, would monitor the current input stream for changes that might signal turn completion: they would (begin to) infer that the end of a turn is forthcoming when the durations of words become longer relative to those of the immediately preceding words and the speaker's speed of speaking decelerates when, crucially, all sorts of other confounds related to word length are controlled for. This hypothesis follows naturally from the on-fly adjustments we know listeners are making to accommodate other characteristics of their interlocutors' input and is, therefore, the hypothesis we will explore here.

In this paper, we will study the speech-rate hypothesis but, given that the (corpus) data we have do not allow us to study listeners' perception and processing, we will focus on the speaker side of it, the degree to which speakers slow down towards the end of turn. That apparent restriction notwithstanding, we will go beyond previous work in several respects, in particular with regard to type and size of the data set to be studied and with regard to the comprehensiveness of the statistical analysis. Obviously, word duration is affected by a number of characteristics including, for example, the phonemic size of words, their frequency, intonation/whether they are the nucleus of a turn, etc. and others, and our statistical analysis will need to be complex enough by involving all these variables as predictors/controls to account for all these things at the same time. The following section will discuss our data set and all the variables – the response, predictors, controls/covariates, and source of random effects – that we are including in our analysis.

2. Data and methods

This research is based on the 'demographically-sampled' subcorpus of the British National Corpus (BNC) featuring natural conversation between intimates and familiars, and those 59 files whose audio files were released in the Audio BNC (Coleman et al., 2012).

Using XQuery (cf. Rühlemann, Bagoutdinov, & O'Donnell, 2015) we extracted from the BNC's conversational subcorpus an initial random sample of 800 10-word turns for which audio recordings were available.[5] The 10-word span was selected as it is likely to represent the average turn length in conversation (Rayson, Leech, & Hodges, 1997, Rühlemann, 2018).

The turns were extracted from the BNC along with metadata including (i) file Ids, (ii) speaker Ids, (iii) position identifiers (slot 1, slot 2, etc.), and (iv) PoS-tags denoting a word form's (likely) part-of-speech in the turn.

Contrary to Local and Walker (2012) above-cited comment that factors such as "syllable and word structure, accentual patterning, position in utterance, speaker, overall speaking rate, information structure etc., are all things which cannot be controlled for and which, moreover, are known to impact on the durational characteristics of words" (Local & Walker, 2012: 259), the regression model we fitted includes, and controls for, precisely the factors Local & Walker mention.

We first provide a brief overview of the variables included in this study and the conceptual roles they play here, before we turn to their exact operationalization:

- the dependent variable: Duration (the acoustic extension of articulation of a word token in a turn in ms);
- the central predictor of interest: PosInTurn (the position of a word[6] token in a turn);
- control variables: Nucleus (most prominently stressed word token in turn); Freq (frequency of word type), Length (phonemic size of word token), Surprisal (informativity of word token in turn), DiffFromMeanPrev (change in mean duration of word tokens in turn);

[4] The list of words occurring at least 5 times in either condition in our sample includes: *'s, but, do, er, he, her, here, i, in, it, know, me, n't, now, off, on, one, out, see, she, that, them, then, there, they, think, this, to, up, was, you*.

[5] The turns selected all met the condition that they contained exactly one <s> unit. These units are used in the BNC for sentence-like, i.e., syntactically and/or pragmatically complete, utterances. While this annotation is not 100% correct, it generally does match turns.

[6] As the data were pulled from the BNC, 'word' is defined here as it is in the BNC: that is, generally, as any alphabetical string between white space. Clitics, however, as well as compounds whose component parts are separated by white space (as in "vice president"), are counted as separate words, whereas orthographically fused compounds such as "greenhouse" are counted as one word.

- random-effect sources of variability: Speaker and File (speaker id in the relevant BNC file) and Word and WordClass (the word itself and its part-of-speech).

To measure Duration as well as Nucleus, the recordings for each turn in the 800 10-word turn sample were accessed through BNCweb, an online interface for the BNC (cf. Hoffmann, Evert, Smith, Lee, & Berglund Prytz, 2008), exported and analyzed in Praat, a phonetic analysis tool (Boersma & Weenink, 2012). Each sound file was listened to repeatedly by a research assistant who was unaware of the research questions of this study; spectral waveforms ('sonograms') were inspected and zoomed in on to determine word boundaries based on 'valleys' in the waveform. Due to poor audio quality, interfering background noises, or distance from the microphone not all 8000 words in the sample could be reliably measured in Praat: also, where word boundaries were blurred, for example due to co-articulation, breathy phonation, unvoiced stops, etc., durations were set to NA. Thus, the number of words whose duration could be ascertained was 7848.

The central predictor, PosInTurn, is straightforward: this is the position of the word in the 10-word turns, i.e. a number from 1 to 10. Less obvious may be the need to factor in a second acoustic measure, namely Nucleus defined as the syllable or word "in a tone unit which carries maximal prominence, usually due to a major pitch change" (Crystal, 2003: 321). The inclusion of Nucleus is crucial on the grounds that it relates to Local and Walker (2012) 'accentual patterns' as well as 'information structure', as there is a "fundamental association between high pitch and new information in English" (Wennerstrom, 2001: 34) with the nucleus marking "the point of information in the sentence that is deemed most valuable or relevant from the speaker's point of view" (Rochemont & Cullicover, 1990: 18; but see Wells & McFarlane 1998 for a critical appraisal of the assumed connection between nucleus and information focus). In coding Nucleus, we recorded the placement of the nucleus in the turn: a nucleus on the first word was coded as 1, a nucleus on the second word as 2, and so forth.

In ascertaining the placement of Nucleus, a conservative approach was adopted to include only turns that featured a single clear nuclear stress, as illustrated in Fig. 1.[7] In Fig. 1, there is a noticeable step-up in pitch of about 3 semitones on the word *optician*, the tenth word in the turn. Thus, for this turn the value 10 was recorded under Nucleus. Any other stress patterns (turns with multiple nuclei, unclear nuclei, or no discernible nuclei at all) were excluded, leading to the loss of 405 turns. The size of the resulting sample – 395 turns – is by no means large but already far larger than the sample sizes used in previous research (s. above).

Several of the controls are fairly straightforward as well: Freq is based on word token counts from the spoken part of the BNC and is included due to the large body of research suggesting that duration is influenced by frequency, with more frequent words being shorter than infrequent words (e.g., Zipf, 1949, Fidelholtz, 1975). Similarly, Length is based on the phonemic length of the word and is included given the trivial fact that words with more phonemes will take longer to articulate. length was computed based on a word list of all the words in the sample; the word types were converted into IPA phonetic transcription using https://tophonetics.com/; words that were not automatically transcribed (e.g., infrequent names or words in non-standard spelling) were manually transcribed; based on these transcriptions, phonemic sizes were established in R.

We also created a control variable Surprisal. This variable is also indicative of what Local and Walker (2012) refer to as 'information structure'. Surprisal, also known as informativity, has been found to correlate with duration, with less surprising words getting reduced in durations (e.g., Seyfarth, 2014) and evidence suggesting that it is "a considerably more important predictor of word length [duration] than frequency" (Piantadosi, Tily, & Gibson, 2011: 3528). The surprisal of an item (Piantadosi et al., 2011, Seyfarth, 2014) such as a word depends on its context. While contextual variables may include, inter alia, discourse, syntax, culture, and world knowledge (Piantadosi et al., 2011) these variables are not easily quantified. We followed the practice of approximating surprisal by computing the negative binary log of the conditional probability of a word given the previous word based on a frequency list of all individual words and file-internal bigrams of the complete spoken component of the British National Corpus (c. 10 million words; cf Hoffmann et al., 2008). Then, two kinds of Surprisal values were computed:

- for each turn's first word, we computed Surprisal as defined in (1);
- for each turn's other words, we computed Surprisal as defined in (2).

$$-log_2 \frac{\text{freq of word in } BNC_{spoken}}{\text{size of } BNC_{spoken}} \tag{1}$$

$$-log_2 \frac{\text{freq of bigram in } BNC_{spoken}}{\text{freq of word}_1 \text{ in } BNC_{spoken}} \tag{2}$$

A quick sanity check revealed that these values 'made sense': very low Surprisal values were found for bigrams such as *gon na*, *ai n't*, *wan na*, [...], *supposed to*, *corned beef*, *thank you*, [...], *willing to*, *kind of*, *pebble mill*, and *couple of*, whereas the highest Surprisal values were found for *her front*, *her wo*, *get passed*, [...], *took Emma*, *your bathroom*, *your keep*, *well something*, [...], *stuff Dave*, *up Sunday*, *he look*, *were because*.

The final control variable is what we call DiffFromMeanPrev. This variable was created from the measurements of the dependent variable Duration. Although based on it, DiffFromMeanPrev differs from Duration in one crucial respect: Duration measurements are only dependent on each word itself and not also on what happened before in the word's turn. DiffFromMeanPrev, by contrast, monitors the changes that Duration undergoes within each turn, i.e. it compares a word's Duration value against the mean of the Duration values of all previous words in the turn. In other words, while all variables including Duration so far capture information 'vertically', that is, across turns, DiffFromMeanPrev captures information 'horizontally', that is, from within turns, allowing us to control for turn-internal information in the regression on Duration. In addition, it also allows us to perform a second regression analysis, one that focuses on within-turn changes by making DiffFromMeanPrev the dependent variable.

[7] We are grateful for constructive feedback on an earlier representation of the pitch contour by John Local.

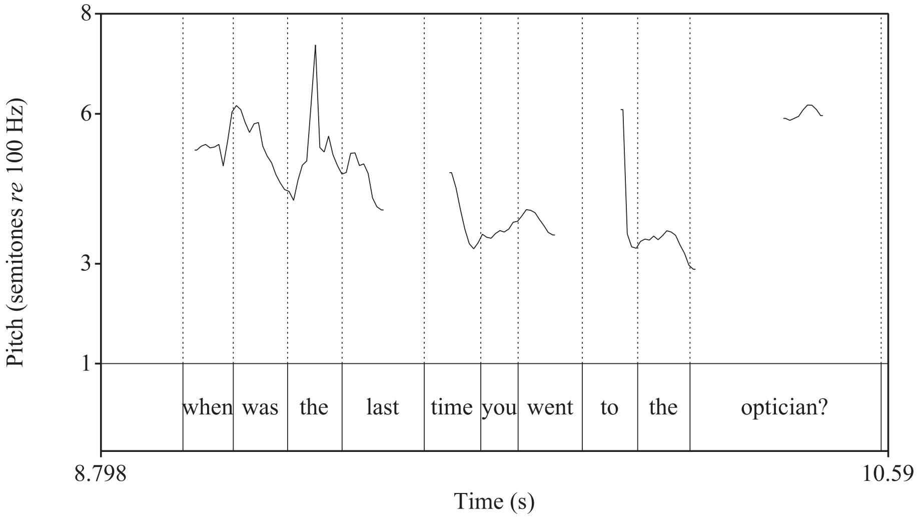

Fig. 1. Pitch contour of turn *When was the last time you went to the optician?*, measured in Praat.

How is the variable DiffFromMeanPrev computed? In order to trace the changes that Duration undergoes within each turn we calculated for each word in a turn but the first, how the duration of that word was different from the mean of all previous word durations in that same turn; the first slot was set to 0. Consider Table 1 for an illustrative example.

The first row contains fictitious Duration data for each word position in the turn; the second row contains the means of all previous Duration values, the final row contains the differences of each Duration value from the mean of the previous ones. For instance, the duration of the third word is 3, the mean duration of all words before that word is $^{1+2}/_{2}$ = 1.5, thus the DiffFromMeanPrev value is 1.5, indicating that that third word is 1.5 units longer in articulation duration than all previous words of this turn. If there is an effect such that Duration changes over the course of a turn, the slope of this control should be significantly different from 0.

As for the random-effects structure of the modeling process to follow, for each word token, we noted the code for speaker who produced it (Speaker, which helps control for speaker idiosyncrasies) and the file in which it occurs (File); obviously, Speaker is nested into File; similarly, we noted the word itself, but also its part-of-speech tag.

We present the analyses in the next section, Section 3: In Section 3.1, we explain the statistical modeling process for the first regression model, in which Duration is the response variable; in Section 3.2, we describe the nature of the effects; finally, in Section 3.3, we come to the results of the second regression model, whose response variable is DiffFromMeanPrev.

3. Statistical analysis and results

In order to prepare the data for the actual statistical analysis, exploration of the data indicated that several preparatory steps were necessary to deal with the usual issues of observational corpus data. First, to address skew, the response variable Duration was transformed into its binary log. The variable Length was logged to the base of two and then *z*-standardized. The variable Freq was square-root transformed (which lead to a slightly more uniform spread of values than the log transformation) and then *z*-standardized; also, Surprisal was *z*-standardized. As for parts of speech, we reduced the variable WordClass to one with fewer levels by retaining only the first three characters from each tag, effectively discarding portmanteau[8] tags in favor of the tagger's first, and preferred, guess.

3.1. The statistical modeling process of Duration

The starting point for our analysis of the hypothesized global turn-final lengthening effect was a model selection process that began with a model that featured:

- Duration as the response variable;
- PosInTurn (polynomial to the 2nd degree to capture curvature) as the central predictor;
- DiffFromMeanPrev, Length, Surprisal, Freq, and Nucleus, but also the interactions of PosInTurn with DiffFromMeanPrev and Nucleus as controls;
- varying intercepts and slopes of PosInTurn for speakers within files and for words and word classes.

This model raised singularity warnings, but we used it only to identify outliers. We computed the residuals of this model and discarded all data points whose absolute residuals exceeded 2.5; this lead to a loss of a mere 1.15% of all data ($^{45}/_{3900}$ data points). We then also simplified the random-effects structure by reducing the nested slopes (speakers within files to just files and parts of speech within words to just parts of speech); since these had very small variances anyway, this proved unproblematic in terms of explanatory power and it took care of the singularity problem.

Next, we tried to reduce the random-effects structure of the model, but both likelihood ratio comparisons and *AICc* comparisons forced us to leave the random-effects structure untouched. Finally, we tried to identify the best fixed-effects structure and arrived at a model that contained curved effects of Length ($p_{\text{LRT/deletion}} < 10^{-15}$), Surprisal ($p_{\text{LRT/deletion}} < 10^{-5}$),

[8] A portmanteau tag indicates ambivalence on the part of the automatic tagger as to what Part-of-Speech tag to assign; this ambivalence is expressed in a hyphenated double tag, e.g., AJ0-AV0, which is used for a word that could be either an adjective or an adverb.

Table 1
Fictitious data to exemplify the computation of DiffFromMeanPrev.

	PosInTurn: 1	PosInTurn: 2	PosInTurn: 3	PosInTurn: 4	PosInTurn: 5
Duration	1	2	3	1	3
Mean of previous durations		1	1.5	2	1.75
DiffFromMeanPrev	0 (per def.)	1	1.5	−1	1.25

Freq ($p_{\text{LRT/deletion}} < 0.0001$), a curved effect of DiffFromMeanPrev ($p_{\text{LRT/DELETION}} < 10^{-15}$), and the interaction of curved effects of PosInTurn and Nucleus ($p_{\text{LRT/deletion}} < 10^{-10}$). This final model has no collinearity issues (all *VIF*s < 1.5), no normality of heteroscedasticity problems in its residuals, and a quite good degree of variance explanation and one that is, fortunately, largely due to the fixed effects: $R^2_{\text{marginal}} = 0.729$ and $R^2_{\text{conditional}} = 0.782$.

As for the interpretation of the model, nearly all of its coefficients are completely uninterpretable because they pertain to orthogonalized polynomials, which is why we will utilize effects plots (Fox, 2003) of predicted Duration values for our discussion.

3.2. The nature of the effects

The first effect to be discussed is one that involves the main predictor of interest, namely the interaction of PosInTurn, which, unsurprisingly given the important role of turn-level stress on duration, interacts with Nucleus. This is represented in Fig. 2 in two ways. The left panel shows a kind of numeric heatmap: the *x*- and the left *y*-axis represent the predictors PosInTurn and Nucleus respectively. The right *y*-axis represents for each value of Nucleus how often it was attested in our data, and the plotted numbers represent the binned predicted duration (with low and high values representing low and high durations).

In the fairly rare situation (≈10% of all cases) when the nucleus of the turn is on one of the first two words of the turn (the bottom area of Fig. 2 when $y \leq 2$), durations actually *de*crease for words later in the turn: The increased length of the nucleus overpowers, so to speak, the rest and does not let the durations develop high values again. However, in nearly all other nucleus positions (i.e., when $y \geq 3$), we see a fairly clear trend such that predicted durations are highest late or very late in the turn; the only exception to this strong trend is the surprisingly high predicted word duration of the first word of the turn when the nucleus is on the last word (see the 8 in the top left corner). In sum, durations decrease continuously for nuclei positioned (very) early in the turn; durations increase continuously for nuclei in mid-turn position; and durations for nuclei positioned (very) late in the turn first decrease toward the middle of the turn before they increase again toward the turn end.[9]

The right panel represents those results in a way that some readers might find more comprehensible. Again, PosInTurn is on the *x*-axis, but now the predicted values of Duration are on the *y*-axes (as is more customary for dependent variables) and, for once, we are zooming into a relatively narrow range of predicted Duration values on the *y*-axes just to make the effect easier to see. This time, every unique value of Nucleus gets its own regression line, which is represented by the respective number. The interpretation does of course not change much but is probably a bit clearer: The regression lines for Nucleus = 1 and Nucleus = 2 go down as PosInTurn increases (as shown in, and discussed for, the left panel), but all other lines nicely swerve up towards the hypothesized higher values of PosInTurn on the right (with different degrees of curvature, obviously), indicating that durations become higher at the end: Speakers are slowing down there.

All remaining effects are those of variables that were only included as controls lest we overestimate the effect of PosInTurn, which is why we do not discuss them in much detail, in particular given that their effects are (reassuringly) as expected; note that, in Figs. 3–6, we are 'zooming out' on the *y*-axis (to be able to represent all observed duration values on the *y*-axes). Fig. 3 is the effect of Freq in the model and, as was to be expected from decades of prior work, more frequent words are pronounced faster and once a certain frequency is reached, that facilitatory effect levels off.

Similarly straightforward is the effect of Length in the model, shown in Fig. 4: Trivially, longer words take longer to articulate and this effect shows no sign of leveling off.

Moreover, consider the control variable Surprisal, whose effect is represented in Fig. 5: More surprising words take longer to articulate. Note that while the regression line seems to suggest that the most surprising words are pronounced more quickly again, this is more likely due to the fact that polynomial effects are affected much less by local trends (and more by the one global trend), meaning that the slight shift downward on the right is most likely *not* due to a speeding-up process there; also, note that, given the very small number of really very surprising words, the confidence band on the right is so wide that this graph is better interpreted as reflecting a slowing down for less predictable words, but one that at some point levels off.

The final effect is that of the last control variable, namely DiffFromMeanPrev, which is represented in Fig. 6: DiffFromMeanPrev is on the *x*-axis, Duration is on the *y*-axis, and the predicted values are represented by the regression line with its 95%-confidence band; the dashed lines are the means of the relevant variables. There is a clear relation such that if a word is longer than the average length of previous words (*x*-axis), its predicted duration increases (*y*-axis). While this is of course not surprising – recall that we included this variable in this analysis as a control – the inclusion of this effect in the model makes sure that any variance in Duration that it picks

[9] As an anonymous reviewer observed, this result is reminiscent of the results obtained in Bögels and Torreira (2015). There, comparing the summed durations of all the words preceding what was the last word in the short question (e.g. "student" in "So you're a student?") but a middle word in the long question ("So you're a student at Radbout University?"), the authors found no significant differences (Bögels & Torreira 2015: 51). At first glance this result seems incompatible with the present result. As the reviewer notes: "[h]owever, looking more closely at Fig. 2, the results actually appear compatible. When the nucleus is at the end (as in the short questions from Bögels & Torreira), the first words are long, but middle words are shorter. However, when the nucleus is in the middle, or near the start, the first words are shorter, but the middle words are a bit longer again. Thus, given that Bögels & Torreira measured the sum of the lengths of all words up to the last one, the end result may have been similar (although the questions were shorter than 10 words)."

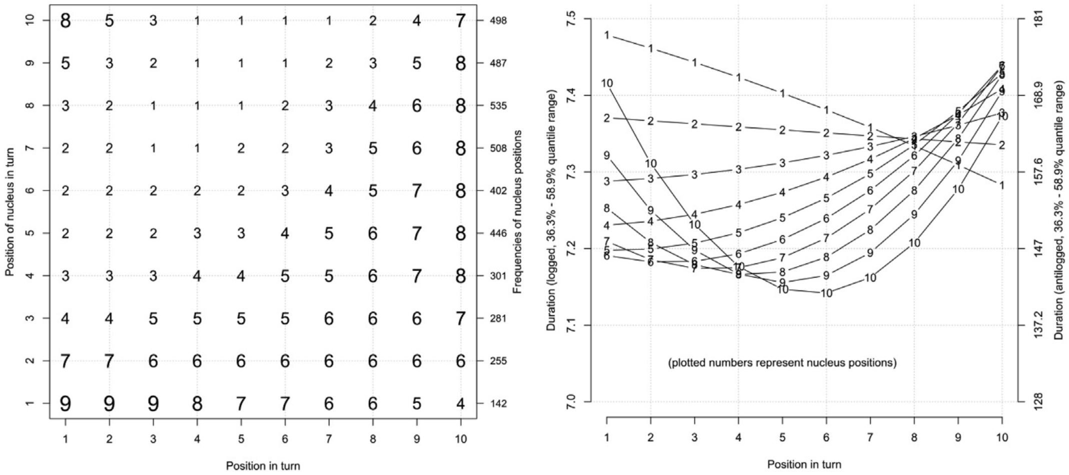

Fig. 2. The effect of PosInTurn and Nucleus in the final model of Duration.

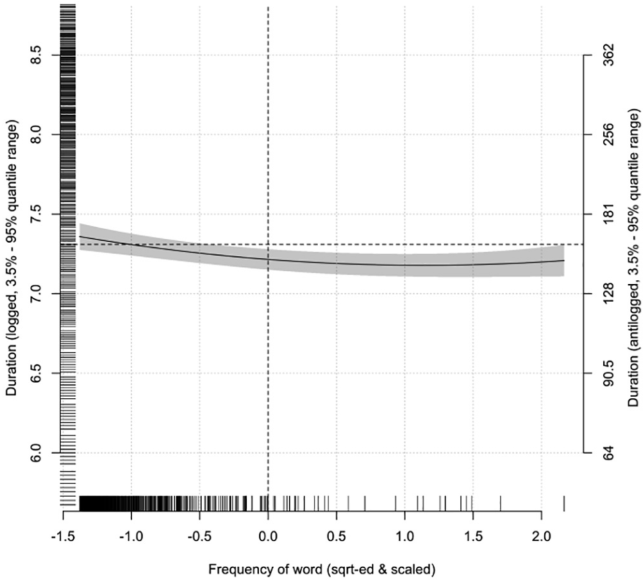

Fig. 3. The effect of Freq in the final regression model of Duration.

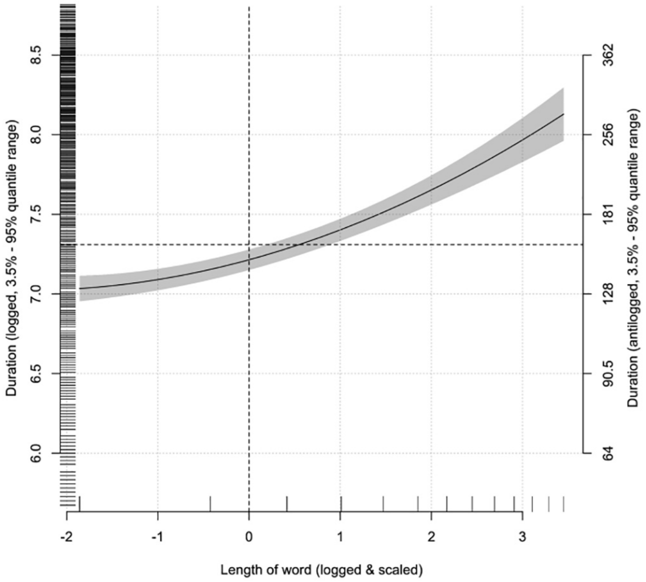

Fig. 4. The effect of Length in the final regression model of Duration.

up on is not anticonservatively attributed to our main predictor of interest, viz. PosInTurn.

3.3. A follow-up model of DiffFromMeanPrev

The above regression modeled Duration, i.e. data points that resulted directly from the Praat measurements of every word token *across turns* showing a clear effect of PosInTurn (mediated by Nucleus) that indicates turn-final lengthening patterns across turns. We now zoom into the changes *within turns* with a regression model whose response variable is DiffFromMeanPrev, which, as discussed above, monitors the changes that Duration undergoes turn-internally. The modeling process was conducted as before, which means that we first trimmed outliers based on residuals of the first model (a mere 1.26% of the data), then we trimmed down the random-effects structure until we arrived at a model without convergence problems (which had varying intercepts for files and parts of speech only), and finally we determined the best fixed-effects structure on the basis of likelihood ratio tests and *AIC*c-comparisons. The final model we arrived at contained curved effects of Length ($p_{\text{LRT/deletion}} < 0.0001$) and Freq ($p_{\text{LRT/deletion}} < 0.0001$), a non-significant control effect of Surprisal ($p_{\text{LRT/deletion}} = 0.33$, a strong effect of Duration ($p_{\text{LRT/deletion}} < 10^{-15}$), and the interaction of curved effects of PosInTurn and Nucleus ($p_{\text{LRT/deletion}} < 10^{-10}$); collinearity and residuals were again unproblematic and the model's fit was quite good: $R^2_{\text{marginal}} = 0.707$ and $R^2_{\text{conditional}} = 0.734$.

In the interest of space, we only discuss the main predictor, PosInTurn, which is again in an interaction with Nucleus and which is shown in Fig. 7.

In the left panel, we have PosInTurn and Nucleus on the *x*- and the *y*-axis respectively and the plotted numbers represent how much the duration of a word in a certain position in a turn changes compared to all previous ones: high(er) numbers mean that a word is (much) longer than the average of the previous words, low(er) numbers mean that a word is (much) shorter than the average of the previous words. Reassuringly,

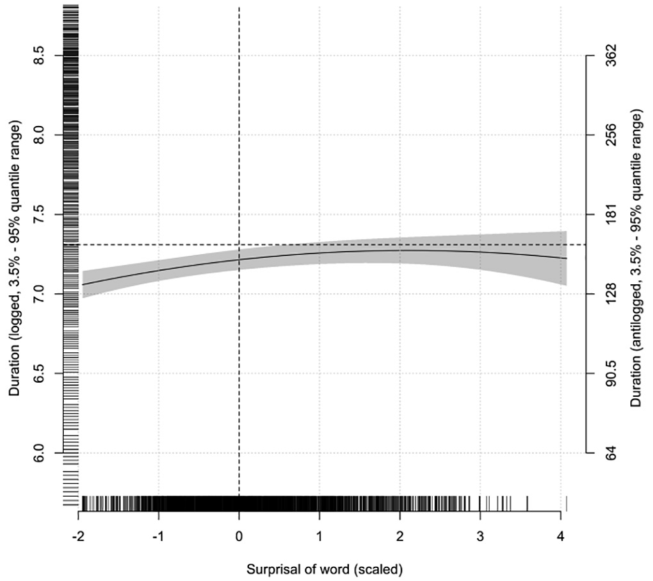

Fig. 5. The effect of Surprisal in the final regression model of Duration.

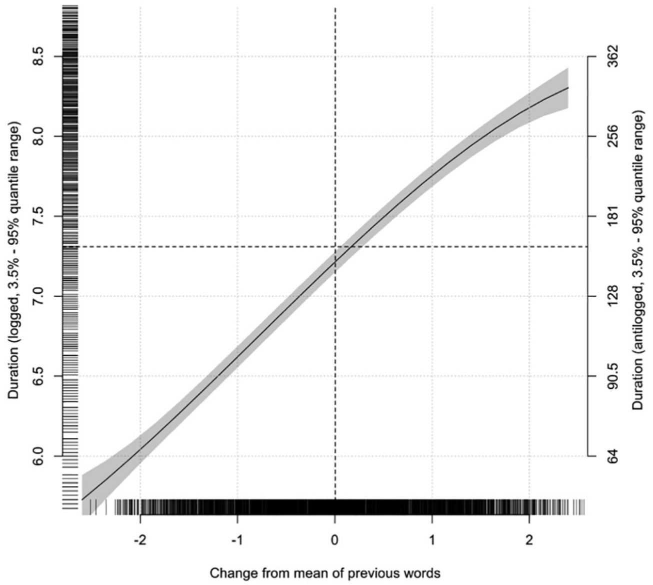

Fig. 6. The effect of DiffFromMeanPrev in the final regression model of Duration.

we see that the highest predictions are made in the top right corner, which is where the turn-final lengthening effect 'meets' nuclei: When the two variables that individually promote long durations come together, we do indeed find the longest words.

In accordance with our expectations, we can also see that words get longer later in the turn when the nucleus is near the end (y > 7). Below that, we see the interaction of two curved trends reflected in the results: When the nucleus is in the middle of the turn (roughly $4 \leq y \leq 7$), then there is little variation: word durations do not change much compared to the previous word durations. However, when the nucleus is at or close to the beginning of the turn ($y \leq 3$), we have a curvilinear effect: the nucleus in one of these early positions is long (er), followed by shorter words that get longer and longer towards the end of the turn.

Again, the right panel shows the same results in a way that might speak to some readers more, but the results are the same: When the nucleus is in positions 8, 9, or 10, the regression lines go up (indicating that the word in each position is predicted to be longer than the previous words' average). On the other hand, when the nucleus is in positions 1, 2, 3 (and to a much lesser extent, 4), the regression lines are *U*-shaped: With the long nucleus at the beginning, words become shorter after that, but then become longer and longer towards the end. Finally, when the nucleus is in positions 5, 6, or 7, the regression lines go down slightly (indicating that the word in each position is predicted to be just a little bit shorter than the previous words' average); we explain this with the assumption that the turn-medial nucleus, with its above average length 'disrupts' any even remotely linear trend one might expect over the course of the turn.

4. Discussion

In this study we have examined what we termed the speech-rate hypothesis, according to which speakers gradually decrease the speed of speaking over the course of the turn to advance-signal turn-completion. We tested this hypothesis using two mixed-effects regression models, one with Duration and one with DiffFromMeanPrev as the dependent variable.

The first model confirmed how subtly word duration reacts to influences from confounds. Not only does a word's duration depend, as is obvious, on its phonemic make-up, that is, whether its phonetic structure is complex or simple, which is an offline property, one the word always has regardless of context. Neither does a word's duration only depend on its overall frequency, yet another offline property, with frequent words being articulated faster than infrequent ones. A word's duration also significantly depends on online factors, that is, factors coming into play once the word is used in discourse and interaction. The model shows that online factors impacting on a word's duration include (i) its surprisal, that is, how much or little the word can be predicted given the context of the preceding word, (ii) nuclear stress, that is, whether the word 'stands out' from the other words in the turn prosodically by higher pitch and energy, and, most importantly in the present connection, (iii) position in the turn, that is, whether the word starts up a turn, falls squarely into the middle of the turn, or concludes the turn. While these factors have been recognized as affecting duration in previous research, this study provided, to the best of our knowledge, the first attempt at modeling these factors and their interactions in comprehensive regression models studying the additional role played by position in turns. Moreover, the models discovered an interaction not, to our knowledge, reported elsewhere: That of nuclear stress and position in the turn. This interaction suggests that when speakers place nuclear stress very early in the turn, specifically in turn-first and turn-second position (which they rarely do) word durations thereafter decrease, indicating that speakers are accelerating the speed of speaking, but when the nucleus falls onto words in any other position in the turn (the much more common scenario) word durations thereafter increase, indicating that speakers are decelerating their speech rate.

In other words, the first model not only allows a rather comprehensive view on a large number of factors impacting on

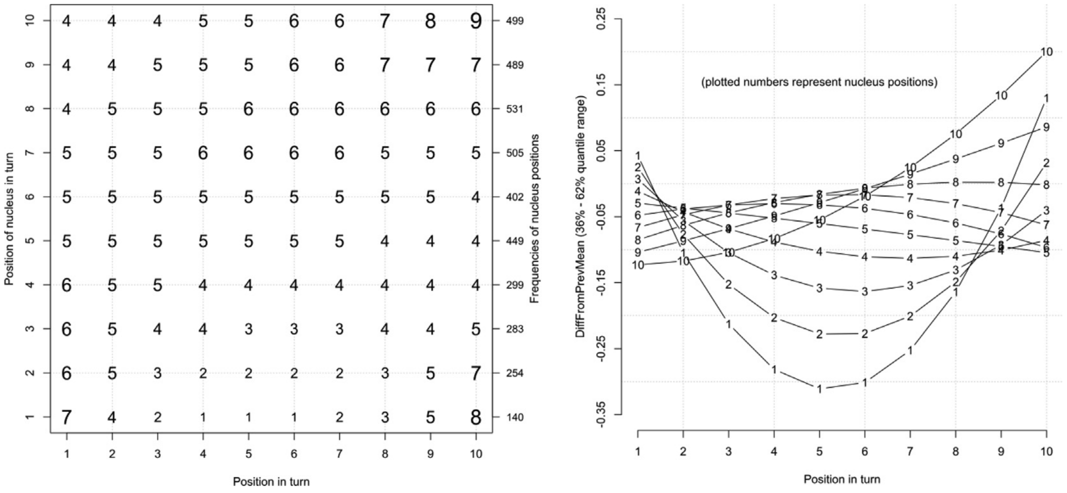

Fig. 7. The effect of PosInTurn and Nucleus in the final model of DiffFromMeanPrev.

word duration in turns-at-talk and teases apart the intricate ways these factors interact; it also provides initial support for the speech-rate hypothesis which we set out to test in this study, according to which turn-final lengthening is really a more global process affecting not only the turn-final word but the turn as a whole.

Compelling evidence to support the speech-rate hypothesis also comes from the second regression model, the one with DiffFromMeanPrev as the dependent variable. To reiterate, the variable DiffFromMeanPrev measured speech rate directly by tracing changes in duration turn-internally, from word to word within one and the same turn. This model suggested that, very much like duration, speech rate depends on the effects of phonemic size, frequency, surprisal, nuclear stress, and position in the turn and is also affected by the interaction between nucleus and position, with the clearest effects on speech rate deceleration shown for turns in which the most heavily stressed word was in turn-final or -prefinal position; the model also indicated that when nuclear stress was placed early in the turn, speech rates thereafter first picked up during the middle of the turn only to return to decelerating tempo toward the end of the turn.

Thus, both models confirm the speech-rate hypothesis: speakers slow down over the course of the turn; however, the slow-down is mediated by nuclear placement. This finding has implications for theories of turn-taking and turn transition. As noted, turn-final lengthening is generally accepted as a cue, out of a large field of cues, by which speakers signal to co-participants the go-ahead to take the turn. Our findings significantly refine this picture: turn-final lengthening is not an isolated phenomenon occurring just on the last word/syllable in the turn; instead, it is the end-point of a process that comprises much larger parts of the turn. This process can be described as *turn rallentando,* that is, as a decrease in the speaker's speed of speaking from turn inception to turn completion. Given the interaction with nuclear placement, the decrease can be continuous and linear, affecting the turn from beginning to end, or discontinuous and curvilinear, affecting only the later parts of the turn. That is, the increases in duration at turn completion points that are observable in the present data and that have been noted in previous research are not the whole signal; they are just the last bit – the culmination point – of the signal. The full signal is the process as a whole, the speaker's slowing-down over the whole turn or its final positions. Either way, the slow-down is not restricted to the last syllable or word in the turn and hence does not mark turn-completion *on its occurrence*. Rather, turn rallentando serves speakers as a resource for *advance-projecting* turn completion very much like morphosyntax, which, as noted in Section 1, provides "long distance projection" (Levinson & Torreira, 2015: 13; cf. also Sacks et al., 1974; Atkinson, 1984; Clayman, 2013; Magyari et al., 2014). As morpho-syntax provides a structural envelope allowing the listener to predict the structural contour of the turn-in-progress, so turn rallentando provides a durational envelope for the listener to predict the durational contour of the turn.

A legitimate question arising from the spread of the signal over (large parts of) the turn is whether listeners can rely on the speaker's rallentando alone in inferring that, and when, the turn is actually complete. Given its being spread out, the decrease in speech rate cannot be seen as pointing out the precise completion point in the turn; rather, the deceleration could continue beyond that point. What the speaker's rallentando, then, does, is indicate that the speaker is *on their way* toward turn completion.

Taking this line of theorizing further, it appears that turn rallentando – and hence the turn-*final* drawl, which is part of it – is insufficient as a stand-alone turn-yielding cue. Rather, turn rallentando is a complimentary cue making turn completion more and more likely but not in and by itself marking out any one point in the turn as the turn-completion point. What turn rallentando indicates – the speaker's notching closer and closer to turn completion – needs to be confirmed yet by some other turn-final cue, potentially one that is not of a continuous but *discrete* nature, be it the return of the speaker's gaze to the listener, the aspiration of word-final plosives, the completion of a 'socio-centric sequence', or the termination of a hand movement, to name only a few (see Section 1), or, most likely, a combination thereof (cf. Bögels & Torreira, 2015: 55).

5. Limitations and future work

As noted in Section 3 above, the fit of the models was good, meaning that most of the variation in the data could be explained by them; given that the data were observational behavioral data, the degrees of variance explanation we achieved just from the fixed effects (72.9% and 70.7%) is quite high. What little variation remained unaccounted for may be explained by the limitations that are placed on the models and the data underlying them.

The first such limitation is the fact that only 10-word turns were examined. While, as noted, the 10-word length is likely the average turn length and hence the central-tendency length in conversation, focusing on a single turn length is an exclusive practice, leaving the considerable variation in turn length – ranging between single-word turns and turns measuring in the hundreds of words (cf. Rühlemann, 2018) – unaccounted for. This is limiting in at least two respects. First, as shown by Yuan, Liberman, and Cieri (2006) turn length impacts speech rate (turns with one to seven words have much slower overall speech rates than, for example, 10-word turns or turns even longer than that).[10] Also, as an anonymous reviewer noted, the exclusive focus on relatively long turns "might have biased the results because it makes it more likely to see graded patterns of changing word length over turns." Shorter turns may not exhibit, or exhibit to a lesser extent, the gradual decrease in speech rate that we observe for 10-word turns; it cannot be ruled out that in, say, 3- or 4-word turns the durational increase is compressed, as it were, on the last word rather than stretched out across (much of) the turn as a whole. An obvious desideratum for future work therefore is to enlarge the pool of turn lengths considered beyond the central-tendency length.

The second, and more consequential, limitation relates to the factors potentially impacting on word duration in discourse and interaction that were not included in the models (although, we should like to add, the number of factors taken into account by our models by far exceed the numbers of factors considered in previous work targeting the role of turn-final lengthening or similar pragmatic factors).

To start with, we only admitted turns with straightforward nucleus placement – those with a single nucleus, thereby evidently excluding the large number of turns (more than half of the turns in the original sample!) where nucleus assignment is less straightforward (cf. Wennerstrom, 2001: 33 ff.). This exclusion is all the more limiting as we have seen that nuclear placement enters, in the context of duration and speech rate, into important interactions with turn position.

Second, the unit of observation for surprisal has been the bigram, the simplest type of context as it merely captures the combination of two adjacent surface forms. While more complex types of context that bear on predictability and informativity, such as culture and world knowledge, will likely remain elusive to quantification for the foreseeable future, the least that can be done in short-term future work is to operationalize surprisal based on trigrams, fourgrams, or even more sizable n-grams; alternatively, predictability and informativity could be measured more broadly using 'overall prior lexical context', a method introduced in Rühlemann and Gries (Forthcoming).

Another limitation is imposed by the fact that the turns in the sample could not be analyzed for internal structure, which, too, might impact on word duration in talk-in-interaction. Trivially, a turn in conversation is a 'turn-in-a-series' (Sacks et al., 1974: 722) interfacing with a prior turn and a next turn. Less trivially, as observed by Sacks et al. (1974: 722), this double interfacing is reflected in the structure of turns, which often have three turn parts consisting of pre-start, turn-constructional unit (TCU), and post-completer.[11] It is particularly this latter turn part that may be relevant in the context of turn-final lengthening and rallentando. The post-completer slot is a "locus of 'current selects next'" (Sacks et al., 1974: 718) and, accordingly, typical occupants of the slot are address terms and question tags. Crucially, post-completers are structures *succeeding* the transition-relevance place (TRP), that is, succeeding the point at which turn transition can 'legally' occur. If the drawl is in fact a signal, or part of a set of signals, by which finishing speakers invite turn transition, the fact that the speaker's turn has already progressed past the TRP would seem to obviate the need for them to extend the drawl into the post-completer. Consequently, post-completers might see no, or reduced, rallentando as the job of flagging the speaker's arrival at the TRP is already accomplished by other means. We are not aware of any research on this matter; thus, this is at present a mere possibility, but one that has some plausibility to it.

Another complicating factor not taken into account is overlap. Numerous turns in the data exhibit turn-final overlap, also referred to as 'terminal overlap' (Jefferson, 1986: 158; cf. also Jefferson, 1973); that is, the speaker has not yet reached turn completion when the next speaker starts up speaking. The commonness in our data is not surprising as overlaps represent the second most common type of transition (after 'slight gaps' of around 200 ms; cf., for example, Stivers et al., 2009).[12] As with post-completers (which often get overlapped, precisely because the TRP has already been reached), it is quite possible that the occurrence of overlap affects how the current speaker signals turn-completion. As shown by French and Walker (1983) for 'turn-competitive overlap', that is, overlap incurred by a next-speaker's early incoming on higher pitch and increased loudness, current speakers, attempting to 'defend' the floor, may resort to increasing intensity and *decreasing speech rate* (French & Local, 1983: 26). Competitive overlap, however, constitutes a marginal portion of all overlap; the lion's share is taken up by 'terminal overlap', which is "massively present" (Jefferson, 1986: 158)[13] and non-competitive as it projects "its almost im- mediate self-liquidation, as the incipiently finishing speaker brings the prior turn to completion" (Schegloff, 2000: 5). Even non-competitive overlaps *may* have an impact in such a

[10] However, our key variable DIFFFROMMEANPREV records changes in speech rate *within* turns not across turns, making it relatively immune to speech rate variation related to turn length variation.

[11] A study by one of the present authors (Rühlemann, forthcoming) specifically investigates forms, functions, and proportions of pre-starts and post-completers in a large sample of turns. The proportion found for post-completers was 15%.

[12] The rates observed for turn transitions occurring in overlap range between 30% (Levinson & Torreira, 2015), 40% (Heldner & Edlund, 2010), and 44% (ten Bosch, Oostdijk, & Boves, 2005).

[13] Clear empirical indication that the overwhelming majority of overlap is non-competitive comes from an analysis of a subsample of randomly sampled 100 between-overlaps and 100 within-overlaps conducted by Levinson and Torreira (2015: 8): they found that 73% of all overlap involved backchannels, that is, short unintrusive response tokens and that obviate the need for special practices for overlap resolution (Schegloff, 2000).

way that speakers no longer ‘bother’ to use any, or the full inventory of, turn-yielding cues as what these are supposed to facilitate – turn transition – has evidently already been achieved.[14]

The final limitation relates to turn function/action and sequence type, factors not included as predictors/controls in the current models. Not every utterance is a full turn. This is true not only of continuer utterances such as “mm” or “u-huh”, which do not constitute a turn but register the listener’s willingness to pass up the turn to the current speaker (cf. Schegloff, 1982). This is also true of those utterances that continuer utterances are a *response to*, namely multi-unit turns, that is, turns built out of multiple turn-constructional units (TCU) (cf. Clayman, 2013: 151), which are typical of ‘telling’ sequences such as instructing, advising, and storytelling (e.g., Schegloff, 2007). Moreover, Torreira and Valtersson (2015: 22) present, and review, compelling cross-linguistic evidence that questions are spoken with faster speech rates than statements, which may affect lengthening/rallentando effects. As with pre-starts and overlaps, due to lack of previous research, we cannot know yet whether speakers’ signaling behavior varies with the action they intend to perform in the turn and the type of sequence they are engaged in. Common sense would again suggest the *possibility* that a speaker in the midst of doing a telling and arriving at one of those “places in it for others’ talk” (Sacks, 1992: 526) may not, or not fully, or not in the same way, signal that arrival at that intermediate point in the sequence as they would signal arrival at the completion point of a less extended turn functioning, say, as a request for information or an offer of assistance and so forth. In future models of turn rallentando, this possibility would have to be controled for.

Notwithstanding these, largely speculative, concerns, we are confident that the discoveries of the present analysis as well as its implications for theories of turn transition will be confirmed even in analyses based on more diverse data and more fine-grained regression models.

6. Conclusions

How co-participants to talk-in-interaction coordinate their conduct to achieve smooth, precision-timed transition from one turn to another is a complex issue as the inventory of resources they draw on to project (as speakers) and predict (as listeners) turn completion is far and wide. This research has taken one such resource into focus: changes in word duration over the course of turns. It has taken head-on the challenges inherent in “[p]roviding robust, quantified, comparative measures of duration” (Local & Walker, 2012: 259) by fitting complex mixed-effects models based on naturally occurring corpus data. Contrary to previous research, which hailed the turn-final drawl as a turn-yielding cue giving non-current speakers the green light to take the turn, the models indicate that drawling, or rallentando, is in fact a far more spread-out process affecting not just the last syllable or word in the turn but large portions of the turn. Rallentando appears to be, not a one-off cue marking the end-point of the turn upon its occurrence, but an extended process advance-projecting the durational envelope of the turn. We have further argued that, as a graded advance-projecting resource, rallentando is in and of itself insufficient to signal turn completion reliably; rather, listeners are likely to rely on turn rallentando *in unison* with other preferably discrete cues marking the turn-completion point upon its occurrence, for “recogniz[ing] that a turn is definitely coming to an end” (Levinson & Torreira, 2015: 12) before triggering the launch of the next turn.

CRediT authorship contribution statement

Christoph Rühlemann: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition. **Stefan Th. Gries:** Methodology, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization.

Acknowledgement

This research was supported by a grant from Philipps-University Marburg.

References

Atkinson, J. M. (1984). Public speaking and audience response: Some techniques for inviting applause. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 370–409). Cambridge: Cambridge University Press.

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society B Biological Sciences, 364*, 1235–1243. https://doi.org/10.1098/rstb.2008.0310.

Barthel, M., Meyer, A. S., & Levinson, S. C. (2017). Next speakers plan their turn early and speak after turn-final “go-signals”. *Frontiers in Psychology, 8*, 393. https://doi.org/10.3389/fpsyg.2017.00393.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication, 52*, 566–580.

ten Bosch, L., Oostdijk, N., & Boves, L. (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication, 47*, 80–86.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics, 52*, 46–57.

Beattie, G., Cutler, A., & Pearson, M. (1982). Why is Mrs. Thatcher interrupted so often? *Nature, 300*, 744–747.

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

Clayman, S. E. (2013). Turn-constructional units and the transition-relevance place. In Jack Sidnell & Tanya Stivers (Eds.), *The handbook of conversation analysis* (pp. 150–166). Malden/MA and Oxford: Wiley Blackwell.

Coleman, J., Baghai-Ravary, L., Pybus, J., & Grau, S. (2012). *Audio BNC: The audio edition of the Spoken British National Corpus*. Phonetics Laboratory: University of Oxford. http://www.phon.ox.ac.uk/AudioBNC.

Crystal, D. (2003). *A dictionary of linguistics and phonetics* (5th edition). Oxford: Blackwell Publishing.

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology, 23*, 283–292.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society, 3*, 161–180.

Duncan, S., & Niederehe, G. (1974). On signalling that it’s your turn to speak. *Journal of Experimental Social Psychology, 10*(3), 234–247.

Fidelholtz, James L. (1975). Word frequency and vowel reduction in English. *Chicago Linguistics Society, 11*, 200–213.

Fox, John (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1–27.

French, P., & Local, J. (1983). Turn-competitive incomings. *Journal of Pragmatics, 7*, 17–31.

Gisladottir, R. S., Bögels, S., & Levinson, S. C. (2018). Oscillatory brain responses reflect anticipation during comprehension of speech acts in spoken dialog. *Frontiers in Human Neuroscience, 12*, 34. https://doi.org/10.3389/fnhum.2018.00034.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language, 25*(3), 601–634.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*, 555–568. https://doi.org/10.1016/j.wocn.2010.08.002.

Hoffmann, S., Evert, S., Smith, N., Lee, D., & Berglund Prytz, Y. (2008). *Corpus linguistics with BNCweb – A practical guide*. Frankfurt am Main: Peter Lang.

Holler, Judith, & Levinson, Stephen C. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences, 23*(8), 639–652.

[14] Alternatively, given the degree to which multi-modal behavior is ‘hard-wired’, speakers’ rallentando might continue unaffected just as deaf signers mouth and even vocalize, and speakers continue gesturing on the telephone (cf. Levinson & Holler, 2014: 6).

Jefferson, G. (1973). A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotics, 9*, 47–96.

Jefferson, G. (1986). Notes on 'latency' in overlap onset. *Human Studies, 9*, 153–183.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22–63. https://doi.org/10.1016/0001-6918(67)90005-4.

Kim, Midam, Horton, William S., & Bradlow, Ann R. (2011). Phonetic accommodation between native and non-native speakers. *Laboratory Phonology, 2*, 125–156.

Kraljic, Tanya, Brennan, Susan E., & Samuel, Arthur G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107*(1), 54–81.

Magyari, L., Bastiaansen, M. C. M., de Ruiter, J. P., & Levinson, S. C. (2014). Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience, 26*(11), 2530–2539.

Local, J., & Walker, G. (2012). How phonetic features project more talk. *Journal of the International Phonetic Association, 42*, 255–280.

Levinson, S. C. (2016). Turn-taking in human communication–Origins and implications for language processing. *Trends in Cognitive Sciences, 20*(1), 6–14.

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B, 369*, 20130302. https://doi.org/10.1098/rstb.2013.0302.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology, 6*, 731. https://doi.org/10.3389/fpsyg.2015.00731.

Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association, 31*(1), 139–152.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America, 108*(9), 3526–3529.

Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics, 2*(1), 133–152.

Rühlemann, C. (2018). *Corpus linguistics for pragmatics*. London/New York: Routledge.

Rühlemann, C. Forthcoming. Turn structure and inserts. *International Journal of Corpus Linguistics*.

Rühlemann, C., Bagoutdinov, A., & O'Donnell, M. B. (2015). Modest XPath and XQuery for corpora: Exploiting deep XML annotation. *ICAME Journal, 39*, 47–84.

Rühlemann, C., & Gries, S. Th. Forthcoming. How do speakers and hearers dismabiguate multi-functional words? The case of well. *Functions of Language*.

Rochemont, M., & Cullicover, P. (1990). *English focus constructions and the theory of grammar*. Cambridge: Cambridge University Press.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language, 82*(3), 515–535.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation'. *Language, 50*(4), 696–735.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen (Ed.), *Georgetown University round table on languages and linguistics analyzing discourse: Text and talk* (pp. 71–93). Washington DC: Georgetown University Press.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society, 29*, 1–63.

Schegloff, E. A. (2007). *Sequence organisation in interaction: A primer in conversation-analysis*. Cambridge: University Press Cambridge.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., de Ruiter, J. P., Yoon, K.-E., & Levinson, C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of the Sciences. U.S.A., 106*(26), 10587–10592. https://doi.org/10.1073/pnas.0903616106.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition, 133*, 140–155.

Torreira, F., & Valtersson, E. (2015). Phonetic and visual cues to questionhood in French conversation. *Phonetica, 72*(1), 20–42.

Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: The time course of response planning in conversation. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.00284.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*(4), 445–472.

Wells, B., & MacFarlane, S. (1998). Prosody as an interactional resource: Turn-projection and overlap. *Language and Speech, 41*(3–4), 265–294.

Wennerstrom, Ann (2001). *The music of everyday speech. Prosody and discourse analysis*. Oxford: Oxford University Press.

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. *Interspeech, 2006*.

Zipf, George K. (1949). *Human behavior and the principle of least effort. An introduction to human ecology*. Cambridge/MA: Addison-Wesley Press.