Speech Reduction, Intensity, and F0 Shape are Cues to Turn-Taking

Oliver Niebuhr
Dept. of General Linguistics
University of Kiel
Kiel, Germany
niebuhr@isfas.uni-kiel.de

Karin Görs
Dept. of General Linguistics
University of Kiel
Kiel, Germany
kgoers@arcor.de

Evelin Graupe
Dept. of General Linguistics
University of Kiel
Kiel, Germany
evelin_graupe@yahoo.de

Abstract

Based on German production data from the 'Kiel Corpus of Spontaneous Speech', we conducted two perception experiments, using an innovative interactive task in which participants gave real oral responses to resynthesized question stimuli. Differences in the time interval between stimulus question and response show that segmental reduction, intensity level, and the shape of the phrase-final rise all function as cues to turn-taking in conversation. Thus, the phonetics of turn-taking goes beyond the traditional triad of duration, voice quality, and F0 level.

1 Introduction

1.1 Empirical background

Signalling turn-taking intentions is essential for successful speech communication. Accordingly, it was shown for all well-described languages that turn holding and yielding cues are robustly encoded in complex redundant bundles of morphosyntactic and/or phonetic patterns. The phonetic patterns primarily rely on prosody, taking a considerable part of its functional load. Compared with turn holding, turn yielding is typically signalled phrase-finally by extensive terminal falling or high rising F0 movements, deviation from modal phonation – mostly in the direction of creak phonation – and increasing final lengthening from penultimate to ultimate syllables. These differences seem to be used in the same way across many languages, and not least for this reason their validity is beyond doubt (e.g., Duncan, 1972; Beattie, 1981; Lehiste, 1982; Kohler, 1983; Nakatani et al., 1995; Ogden, 2001; Fon, 2002; Kohler, 2004; Peters, 2006; Vaissière & Michaud 2006; Gravano, 2009; Fon et al., 2011).

However, leaving aside gaze and gesture patterns (cf. Kendon, 1995; Taboada, 2006), a growing body of evidence from production studies suggests that the phonetics of turn-taking is still richer and goes beyond the traditional triad of voice quality, duration, and the level or direction of F0 patterns. Turn holding or yielding also seems to include the fourth prosodic dimension, i.e. *intensity*, as well as details in the *shape* of phrase-final rises and the degree of phrase-final *segmental reduction*.

For example, phrase-final voiceless plosives in English are realized either fully pronounced and with strong post-aspiration, or in reduced forms that lack post-aspiration and are (partly) replaced by glottalization, cf. "got" [gʷɒtʰ] vs. [gʷɒʔᵗ], and "cap" [kʰæpʰ] vs. [kʰæʔpˀ]. The difference between the unreduced and reduced forms was for a long time claimed to be a matter of free variation, until it was revealed in corpus analyses of different varieties of English that reduced forms were produced turn-medially whereas unreduced forms occurred almost exceptionally at the end of a speaker's turn (Local et al., 1986; Docherty et al., 1997; Local & Walker, 2012).

More recently, it was additionally found for English and French in independent analyses of spontaneous speech corpora that the intensity levels of phrase-final syllables differ depending on whether the syllables occur turn-medially or turn-finally (Gravano & Hirschberg, 2009; Clemens & Dieckhaus, 2009; Raux, 2008; Friedberg, 2011). The difference was the same in both languages: "*speakers tend to lower their voices when approaching turn boundaries, whereas they reach turn-internal pauses with a higher intensity"* (Gravano & Hirschberg, 2009:256).

Furthermore, Dombrowski & Niebuhr (2005) showed on the basis of one of the largest corpora of Standard Northern German – the Kiel Corpus of Spontaneous Speech – that it is not only the range of phrase-final intonation movements that

Proceedings of the SIGDIAL 2013 Conference, pages 261–269,
Metz, France, 22-24 August 2013.

distinguishes turn-internal from turn-final boundaries. At least in the case of rises, it also matters whether the shape of the rise is concave (slow rise followed by fast rise) or convex (fast rise followed by slow rise). Convex rises occurred predominantly turn-medially, whereas concave rises were used by speakers almost invariably at the end of a turn. A similar distribution of rise shapes was found by Asu (2006) for discourse markers in spontaneous dialogues of Estonian.

1.2 Question and aim

The three groups of cross-linguistic findings on reduction levels, intensity levels and rise shapes have in common that their perceptual relevance for turn-taking has never been tested as yet. That is, do listeners actually interpret phrase-final differences in (i) the degree of segmental reduction, (ii) the intensity level, and (iii) the shape of F0 rises as signals of turn-holding and/or turn-yielding? Providing a first answer to this question is the main aim of the present paper.

Clayards et al. (2007) showed that the more systematically acoustic cues are used in speech production the more likely they are exploited by listeners. Given the distinct production findings for (i)-(iii) and their consistency across languages or language varieties, it was already expectable that the answer to our question would be 'yes'; and, indeed our results met our expectation. Yet, empirical testing was indispensable.

1.3 Research subject

Our study was based on a single language variety: Standard Northern German. However, in view of the strong cross-linguistic parallels in the phonetics of turn-taking (cf. 1.1), it is reasonable to assume that our results will also be applicable to many other languages.

In order to test the effects of intensity differences and particularly of reduction differences on the perception of turn-internal and turn-final boundaries, we used the most frequent sonorous word-final syllable in German: <-en#>. It always occurs unstressed and is phonologically represented as a sequence of schwa and alveolar nasal /ən/. However, next to its corresponding canonical pronunciation as [ən] (or rather [n̩]), the word-final <-en#> syllable is known to undergo different reduction processes. The two most important processes are /ə/ elision, which leaves a syllabic nasal, and assimilation of the syllabic nasal to the place of articulation of the preceding consonant. For example, "lieben" (to love) can be realized as [ˈliːbən], [ˈliːbn̩], or [ˈliːbm̩]. Likewise, possible pronunciations of "sagen" (to say) are [ˈzaːgən], [ˈzaːgn̩], and [ˈzaːgŋ̩].

However, prior to conducting any perception experiments, we first had to confirm that the differences in the turn-internal vs. turn-final reduction and intensity levels found for English and French (cf. 1.1) do occur as well in Standard Northern German (the differences in rise shape are already known for Standard Northern German and thus need not be replicated). Therefore, our perception experiments were preceded by an analysis of the Kiel Corpus of Spontaneous Speech. This analysis is detailed below.

2 Corpus analysis

2.1 Analysis method

The Kiel Corpus of Spontaneous Speech includes 117 dialogues which add up to more than four hours of Standard Northern German speech from 52 male or female interlocutors (Kohler, 1996). The corpus is completely annotated, segmentally and prosodically. The segmental annotations are made such that they specify reduction processes like assimilation, elision, lenition, and "articulatory prosodies" in terms of deviations from the canonical forms of the spoken words (articulatory prosodies preserve the "phonetic essence" of segmentally elided words or syllables in the form of suprasegmental sound qualities and are thus an important cue to word identification in reduced speech, Kohler & Niebuhr 2011). Furthermore, the structure of the corpus in combination with the prosodic annotation allows a differential search for phrase and turn boundaries.

On this basis, we conducted an annotation-based analysis of unstressed word-final <-en#> syllables in turn-final and turn-internal position. The turn-internal tokens were further subdivided into phrase-final and phrase-internal syllables. The latter phrase-internal syllables were not directly relevant for our research question but still included for the sake of completeness. Our corpus query yielded a total of 17,023 word-final <-en#> syllables. The majority of the syllables, viz. 11,329 tokens, occurred in phrase- and turn-internal position. For the phrase-final but turn-internal position, we found 4,090 <-en#> syllables. The phrase- and turn-final position was represented by 1,604 tokens. The information about whether the <-en#> syllables were subject of reduction processes, and if so, whether degree of reduction differed across the three syntactic-prosodic positions was derived from the segmen-

tal annotation. We focussed on the two main reduction processes exemplified in 1.3: /ə/ elision and, if the /ə/ is absent, additional progressive place assimilation of the syllabic nasal /n/ towards [m] or [ŋ].

In a subsequent step, we took random subsamples of 50 <-en#> syllables from each of the three syntactic-prosodic positions and analyzed their intensity levels manually in Adobe Audition. Measurements were taken in terms of mean acoustic energy (in dB). As mixing syllables with and without schwa could have biased our intensity measurements, all three sub-samples only contained syllabic [n̩] nasals. The results of our reduction and intensity analyses are presented in the following section.

2.2 Results of the production data

To put it in a nutshell, analyzing the segmental annotation of the Kiel Corpus clearly showed: The more finally a <-en#> syllable was produced the lower was its degree of segmental reduction. This fact is illustrated in Figures 1(a)-(b). While virtually no <-en#> syllable in turn-medial *and* phrase-medial position was realized with a [ə] or a similar vocoid sound before the nasal, about 7% of the <-en#> syllables in turn-medial *but* phrase-*final* position showed such a vocoid section (Fig.1a). The amount of schwas or similar vocoids increased above 10% for those <-en#> syllables that occurred phrase-finally *and* turn-finally. Among the <-en#> syllables that were realized without /ə/, the frequency of place assimilation of the syllabic /n/ decreased from almost 80% in phrase-medial and turn-medial position, through about 66% in turn-medial *but* phrase-*final* position, to only about 20% in phrase-final *and* turn-final position (Fig.1b).

Although these figures speak for themselves, we also assessed their statistical significance by means of a χ^2 test. The test was based on the absolute number of /ə/ and /n/ occurrences in the 3x2 conditions of Figures 1(a)-(b). The test statistics corroborate that reduction becomes significantly stronger under increasing finality (χ^2=373.554, *df*=2, p<0.001).

A similar tripartite picture emerged for the intensity measurements. The intensity level in the random sub-samples of 3x50 <-en#> syllables (realized as syllabic nasals) decreased successively by on average about 3.5-6.2 dB (for female speakers less than for male speakers) from phrase- and turn-medial tokens to tokens which are both phrase-final *and* turn-final. That is, the softest <-en#> syllables occurred immediately before a turn transition. A one-way ANOVA showed that the intensity decrease across the three finality conditions was highly significant (F[2,147]=45.941, p<0.001).

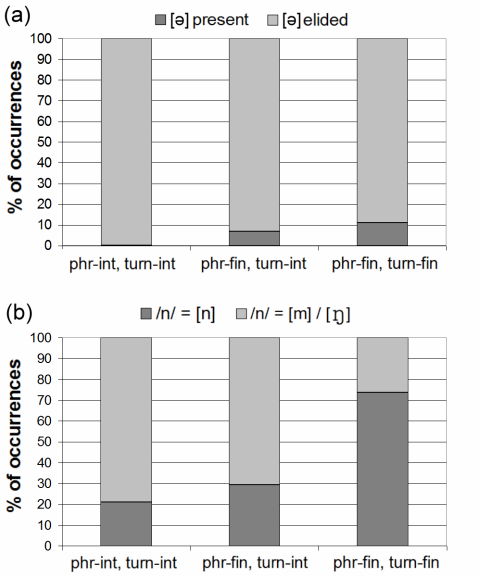

Figure 1: Degree of reduction of <-en#> syllables in terms of (a) /ə/ elision and (b) place assimilation of /n/ by the preceding consonant (when /ə/ is absent). The <-en#> syllables occurred in phrase- and turn-internal position (left), in turn-internal but phrase-final position (middle), or in phrase- and turn-final position (right).

2.3 Conclusion from the production data

Three conclusions can be drawn from the results of our corpus analysis. First, the degree of reduction of sound segments in Standard Northern German – represented by <-en#> syllables – differs substantially depending on whether they coincide with phrase boundaries or turn boundaries. The degree of reduction is lower at turn-final than at turn-internal boundaries. Second, also the intensity levels before different types of syntactic-prosodic boundaries show clear differences. However, while the degree of segmental reduction *de*creases from phrase-internal through phrase-final to turn-final boundaries, the degree of intensity reduction *in*creases by on average up to 100%. Third, the findings for Standard Northern German, particularly the direction of changes from phrase-*final* but turn-*internal* to phrase- *and* turn-final boundaries, are qualitatively consistent with those that have been found before for spontaneous dialogues in English and French.

After having confirmed that Standard Northern German resembles English and French with

regard to the production of reduction and intensity differences at turn-internal and turn-final boundaries, we continued with conducting two perception experiments. They were based on question stimuli, whose ends were varied in a binary fashion with respect to <-en#> reduction, intensity level and the shape of the final F0 rise.

3 Perception experiment 1: reduction and rise shape

3.1 Stimulus generation

When it comes to testing the perceptual relevance of phonetic details for turn yielding or holding, internal/ecological validity is a big issue. We addressed this issue by developing an interactive experimental design in which the participants gave real verbal responses to the stimuli. Our target stimuli were syntactically marked questions, whose last constituent was concluded by a target word. As there were 16 different questions, we had 16 different target words. All of them were similarly frequent verbal infinitives of two or three syllables, and with lexical stress and a rising nuclear pitch accent (L+H*) on the penultimate syllable. The pitch-accent rise was complemented by a high boundary tone (H-%), and hence the rise continued across the final syllable until the end of the utterance. The final syllable was <-en#>. Two target-word examples have already been given in 1.3; further examples are “fliegen” (to fly), “liegen” (to lie), “kramen” (to fish sth out), and “fragen” (to question). In half (i.e. eight) of the target words, <-en#> was preceded by a labial consonant (/m,b/). The other half had a velar consonant (/ŋ,g/) before the <-en#> syllable. Moreover, the target words were balanced with respect to vowel quantity and height of the stressed vowel (/i(:)/ or /a(:)/).

The target questions were embedded in context frames, i.e. they were preceded by 1-2 introductory statements and followed by an alternative question starting with “oder” (or). For example, “Ich hab Anjas Freund letztens Hand in Hand mit einer anderen durch die Stadt laufen sehen. Meinst Du, ich soll Ihr das sagen ? – oder soll ich mich da lieber raus halten?“ (I saw Anja‘s boyfriend yesterday wandering hand in hand through the streets with another girl. Do you think I should tell her? – or should I rather butt out?). The crucial point is that the alternative question is optional. It may or may not be there so that the target question could equally be turn-internal or turn-final. In order to validate this positional ambiguity, we ran a pretest with 12 participants and orthographic representations of our target questions. The pretest confirmed that none of the target questions had a semantic bias towards occurring in turn-internal or turn-final position (i.e. with/without an alternative question).

The 16 sequences of preparatory statement(s), target question and alternative question were produced by a phonetically trained female speaker (KG) with unreduced, canonically pronounced <-en#> syllables ([ən]) at the end of the target words. The sequences were digitally recorded and constituted our first set of 16 base stimuli. Then, KG produced the same 16 sequences again, but this time the <-en#> syllables were highly reduced to either [m̩] or [ŋ̍]. The latter segments were used to create a second set of 16 base stimuli by taking the stimuli of the first set and replacing (with Adobe Audition) their fully pronounced [ən] syllables with the corresponding highly reduced nasal. In this way, we ended up with two sets 16 base stimuli. The stimuli in each set were phonetically absolutely identical except for the <-en#> syllables, which were either fully pronounced or highly reduced.

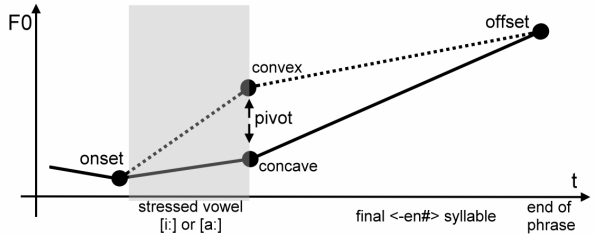

Figure 2: Shape manipulation of the nuclear F0 rise L+H* H-% at the end of the target questions, yielding 32 questions with convex and 32 questions with concave rises.

Before proceeding with the next step, we checked our base stimulus endings for confounding turn-taking factors. First, there were no deviations from modal phonation in the target syllables. Second, final lengthening was also controlled insofar as the fully pronounced and highly reduced target syllables showed no systematic duration differences. Moreover, all duration differences were below the just noticeable difference of 20% (cf. Klatt 1976), which corresponds to about 40 ms in the case of our target syllables.

All 32 base stimuli were then subjected to a PSOLA manipulation in PRAAT, in which we firstly set the pause between target and alternative question to exactly 1.5 seconds. For reasons that will be explained in 3.2, this interval is at the upper limit of turn-internal pauses in dialogues and thus suitable to raise the reasonable suspicion that no alternative question would follow

(e.g., Edlund & Heldner, 2005). Finally, we replaced the naturally produced F0 patterns at the end of the 32 target questions (all of them were more or less linear rises) by clearly convex and concave rises of the same overall range, as is illustrated in Figure 2. The rises were stylized at rise onset, pivot, and rise offset. They set in right before the stressed vowel; the pivot was located at the end of the stressed vowel.

The PSOLA manipulations resulted in 64 resynthesized stimuli. Another 64 stimuli were created by cutting off the alternative questions from the 64 resynthesized stimuli.

3.2 Subjects and procedure

Twenty native speakers of Standard Northern German participated in the perception experiment (14 females, 6 males, 20-30 years old). All participants were undergraduate students of Empirical Linguistics at the University of Kiel.

The participants sat in a sound-treated room and put on a headset. Then, they were instructed that they would be presented with 64 stimuli, each of which would end in a question. Their task would be to conceive themselves in a dialogue situation and to respond to the questions of their female dialogue partner with short, plain answers ('yes/no', 'don't know', 'we'll see' and the like) as soon as they would think that they were given the floor. However, if they answered too early, i.e. before their dialogue partner's turn had ended, their answer would count as a failure. On the other hand, if they did not respond within 1.5 seconds after their dialogue partner's turn had ended (indicated by a bleep), then their answer would count as a failure, too. At the end of the experiment, that participant who gave the most valid answers in the shortest average response time would win a prize (a 50 € voucher).

The crucial point in this procedure was that the participants did not know when the target question was turn-internal or turn-final, i.e. when it was followed by an additional alternative question, as this variable was randomly distributed across the 64 stimuli. In this way, we avoided that the participants were able to learn artificial turn yielding or holding cues during the experiment by correlating the phonetic variation at the end of the target questions with the occurrence of alternative questions. Furthermore, informal pretests showed that the dichotomous forces of the competitive task – i.e. the risk of premature vs. overdue responses – were effective in making participants focus on the stimuli and exploiting given acoustic cues.

Prior to the actual experimental session, which took about 20 min, the participants received a practice session with 12 stimuli that were randomly selected for each participant. The 64 stimuli of the subsequent experimental session were also played in individually randomized orders.

The entire experimental sessions of all participants were recorded via their headsets. Recordings were made with Audacity in the form of digital stereo files with separate channels for stimuli and responses. On this basis, we measured the response times, i.e. the time intervals from the end of the target question to the first response signal of the participant (which included smacks). This response-time measure (in ms) served as dependent variable. Response-time measurements were made manually in Audacity. Figure 3 displays an example. If the relevant first response came too late (e.g., after an alternative question had begun) or not at all, response time measurements were capped at 1.5 seconds.

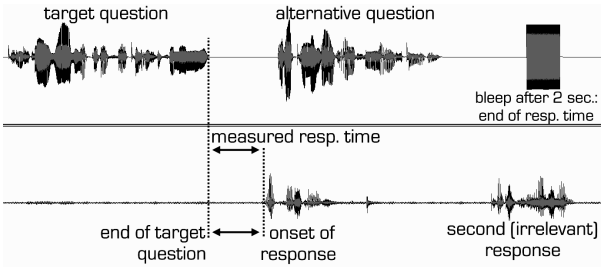

Figure 3: Audacity screenshot showing an example of a response-time measurement; top: stimulus channel, bottom: response channel.

3.3 Hypotheses of Experiment 1

Assuming that participants would respond more readily/reluctantly when they perceived turn-yielding/-holding cues in the target questions of their virtual female dialogue partner, we hypothesized that response times would be shorter (i) for target questions ending in concave than in convex rises and (ii) for target questions ending in unreduced [ən] syllables rather than in the reduced syllabic nasals [m̩] or [ŋ̍].

Although we used target questions with different wordings (that included the target words), a pretest showed that none of these wordings created a semantic bias towards a turn-internal or turn-final interpretation (cf. 3.1). Thus, we expected no effect of the variable Question Wording on response times. The same was true for the target-word internal variable Segmental Context (<-en#> preceded by a labial or velar consonant).

3.4 Results of Experiment 1

The results of the first perception experiment are depicted in Figure 4 in terms of response times per stimulus condition, averaged across all 20 participants. For the statistical analysis, we used a four-way repeated-measures ANOVA (n=20) based on the fixed factors Reduction (2 levels), Rise Shape (2 levels), Segmental Context (2 levels), and Question Wording (8 levels). The ANOVA yielded three significant main effects on the dependent variable Response Time (in ms). The main effects concerned Reduction ($F_{[1,19]}$= 57.716, $p<0.001$, η_p^2= 0.752), Rise Shape ($F_{[1,19]}$= 63.462, $p<0.001$, η_p^2= 0.770), and Segmental Context ($F_{[1,19]}$= 10.991, $p<0.001$, η_p^2= 0.366). The factor Question Wording was not significant, neither was any of the interactions. Insofar, our results allow a straightforward analysis. As is shown in Figure 4, response times were significantly shorter …

- when the target questions ended in the unreduced [ən] syllables rather than in the reduced syllables [m̩] or [ŋ];
- when the rising intonation at the end of the target questions had a concave rather than a convex shape, i.e. when the F0 rise started shallowly across the initial, accented syllable of the utterance-final verb and continued steeply until the end of the utterance (cf. Fig.2);
- when the final <-en#> syllable was preceded by a labial rather than a velar plosive or nasal.

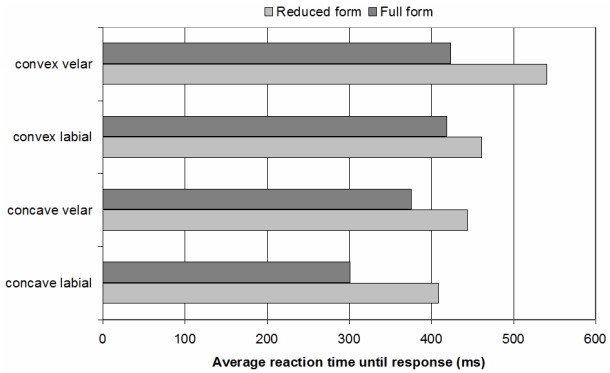

Figure 4: Results of Experiment 1 in terms of average response times (in ms) per stimulus condition; each bar n=20.

3.5 Conclusions from Experiment 1

Our hypotheses concerning the effects of rise shape and the degree of (question-)final reduction on the participants' response times were confirmed. Participants responded slowest after target questions that ended in a convex final rise across a widely reduced <-en#> syllable, and they responded fastest after target questions that ended in a concave final rise across an unreduced <-en#> syllable. The latter target questions caused response times of only 300-400 ms, which is exactly in the order of magnitude of successful – i.e. intended and correctly interpreted – turn transitions in German (Weilhammer & Rabold, 2003). This fact lends further support to our main conclusion: Rise shape and degree of reduction, which were both found to vary utterance-finally on the part of speech production, also function as cues to turn-yielding and/or turn-holding on the part of speech perception.

Moreover, as expected on the basis of our semantic pretest, Question Wording had no effect on response times. However, we found an unexpected response-time effect of Segmental Context, i.e. the place of articulation of the consonant that preceded the <-en#> syllables. We have no clear explanation for this finding as yet. It could be an experimental artefact caused by different frequencies of occurrence of our 'labial' and 'velar' target words. Such frequency differences, even if they are small, could be associated with different reduction baselines. These baselines could have then interacted differently with the turn-taking interpretation of our Reduction variable. Alternatively, the effect of Segmental Context could be due to a difference in intrinsic intensity, which is slightly higher for velar than for labial consonants. This difference also applied to the final nasal /n/ when it was assimilated to [m̩] or [ŋ]. That a lower/higher intensity level towards the end of utterances can basically be interpreted as a turn-yielding and/or turn-holding cue will be shown in the following Experiment 2.

4 Perception experiment 2: reduction and intensity level

Although Experiment 2 primarily tested, if and how utterance-final intensity variation affected the listeners' response times, it was additionally used to investigate the reduction effect of Experiment 1 in more detail. In Experiment 1, we contrasted widely reduced <-en#> syllables ([m̩], [ŋ]) with their maximally unreduced counterpart [ən], being aware of the fact that [ən] is a rare realization of <-en#>, cf. Figure 1(a). Now, in Experiment 2, we turned to the much more frequent, but also perceptually much more subtle reduction difference in <-en#> syllables: assimilated and non-assimilated syllabic nasals.

4.1 Hypotheses of Experiment 2

Our hypotheses were that response times would be shorter (i) for target questions ending in low rather than in high intensity levels and (ii) for target questions ending in less reduced (non-assimilated) [n̩] rather than in reduced (assimilated) [m̩] or [ŋ̍]. In addition, we expect to replicate the secondary findings of Experiment 1: There should be no systematic effect of Question Wording; and, concerning the Segmental Context, there should be faster response times for the group of target words with 'labial' consonants.

4.2 Method of Experiment 2

The method was the same as in Experiment 1, except for three points.

First, we performed a second round of recordings in which we attenuated the reduction difference at the end of the target questions from [ən] vs. [m̩]/[ŋ̍] to [n̩] vs. [m̩]/[ŋ̍]. That is, all <-en#> syllables were realized as syllabic nasals so that the reduction difference became only a matter of presence vs. absence of place assimilation by the preceding labial or velar consonant. The pause between target and alternative question in the stimuli was again set to 1.5 seconds by inserting or cutting out silence.

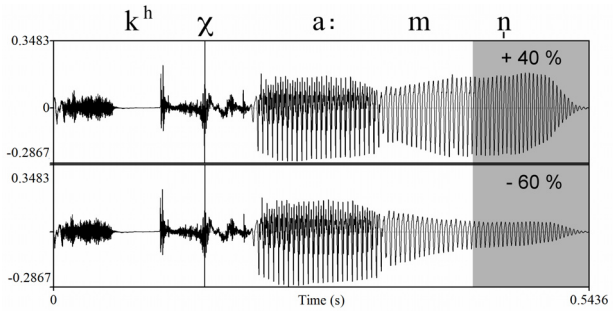

Figure 5: Intensity contrast (loud: +40%, top vs. soft: -60%, bottom) created for Experiment 2, exemplified by "kramen" (to fish sth. out).

Second, instead of the difference in rise shape (all target questions of Experiment 2 ended in a similarly concave rise), we created a difference in the intensity level of the <-en#> syllable. More specifically, the naturally produced intensity pattern of each syllabic nasal [n̩], [m̩] or [ŋ̍] (which typically featured a small intensity increase towards the center of the sound segment, followed by a steep intensity decrease until the end of the target question) was expanded by 40% for the "loud" condition and abated by 60% for the "soft" condition. So, the resulting pairs of stimuli showed a clearly perceivable intensity contrast – "loud" vs. "soft" – in the amount of 100% or 6 dB at the end of the target questions. The intensity manipulation was conducted with Adobe Audition. An example of two waveforms of the question-final target word "kramen" (to fish sth. out, without place assimilation of /n/) is presented in Figure 5.

Third, Experiment 2 was run with a different group of 20 native speakers of Standard Northern German (15 females, 5 males, 23-38 years old).

4.3 Results of Experiment 2

We used again a four-way repeated-measures ANOVA for analyzing our response-time measurements. The fixed factors were the same as in Experiment 1, except that the former factor Rise Shape was substituted by the factor Intensity Decrease. The results of the ANOVA are restricted to three significant main effects that concerned the fixed factors Reduction ($F_{[1,19]}$= 324.653, $p<0.001$, η_p^2= 0.945), Intensity Decrease ($F_{[1,19]}$= 460.355, $p<0.001$, η_p^2= 0.96), and Segmental Context ($F_{[1,19]}$= 72.091, $p<0.001$, η_p^2= 0.791).

As can be seen in Figure 6, the effect of Reduction is due to the fact that participants responded more quickly in the less reduced [n̩] condition than in the reduced [m̩] or [ŋ̍] conditions. Furthermore, responses came faster when the degree of intensity reduction at the end of the target question was stronger, i.e. when target questions ended softer rather than louder. Finally, response times were shorter when the syllabic nasal at the end of the target question was labial rather than velar and/or preceded by a labial rather than a velar consonant.

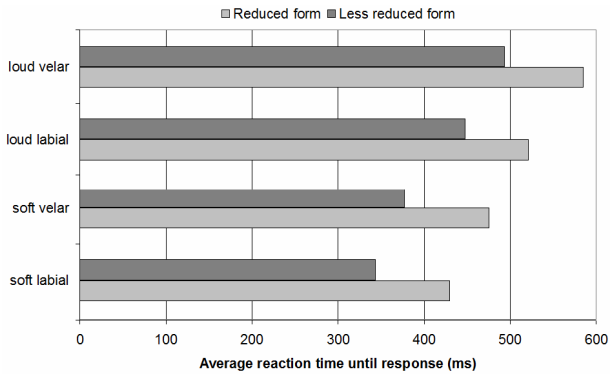

Figure 6: Results of Experiment 2 in terms of average response times (in ms) per stimulus condition; each bar n=20.

4.4 Conclusions from Experiment 2

All hypotheses in 4.1 were supported by Experiment 2 The effect of Reduction means that our participants detected the subtle phonetic differ-

ence between non-assimilated and assimilated question-final /n/, and their reaction to this subtle difference was consistent with that of Experiment 1: Participants respond readily when the target question ended less reduced, and/or they hesitated to respond after those target questions whose final <-en#> syllables were more strongly reduced. So, Experiment 2 provided additional evidence for our conclusion that stronger segmental reduction functions as a cue to turn-holding and/or that weaker segmental reduction functions as a cue to turn-yielding.

The intensity level of the final <-en#> syllables had a separate effect. That louder ending questions delayed the participants' response times suggests that a high utterance-final intensity level has a turn-holding function. Additionally (or alternatively), the immediate responses after soft ending questions indicate that the lower utterance-final intensity level is a cue to turn-yielding. Like for the segmental effects above, the intensity effects fit in well with the real use and distribution of intensity differences in spontaneous dialogues, cf. 1.1.

If we view the loud vs. soft contrast in terms of a low vs. high degree of reduction, then we can see that intensity reduction and segmental reduction played opposite roles in turn-taking. Stronger reduction at the segmental level pointed in the direction of turn-holding, whereas stronger intensity reduction pointed in the direction of turn-yielding. This fact leads to the conclusion that turn-taking cues are no indexical cues insofar as they cannot be uniformly projected onto changes in the speaker's effort.

Furthermore, Experiment 2 also replicated the Segmental-Context effect of Experiment 1. Assuming – in accord with previous studies and informal measurements in our stimuli – that the labial condition was associated with an overall lower intensity level in the target words than the velar condition (e.g., due to longer closure durations, less intense releases, and closed lips during nasal production), then the unexpected Segmental-Context effect becomes understandable as an additional reflection of the role of intensity in turn-taking. That is, as has been anticipated in 3.5, the intrinsically higher intensity in our velar target words created a bias towards turn-holding, and/or the intrinsically less intense labial target words created a bias towards turn-yielding.

Finally, a comparison of Figures 4 and 6 shows that the response times yielded by Experiment 2 were overall longer than those of Experiment 1. This general bias should not be over interpreted as it is probably just due to the fact that Experiment 2 was performed in the evening, whereas Experiment 1 took place in the morning.

5 General discussion

It is known for a long time that turn yielding and holding rely on complex form-function systems. So far, these systems have been typically associated with the prosodic triplet of duration, voice quality, and F0 level. Together they create bundles of perceptually salient phrase-final patterns that involve the direction and range of intonation movements, final lengthening, and non-modal voice qualities (typically glottalization).

More recently, analyses of spontaneous dialogues suggested that the bundles of final turn yielding and holding cues are still richer and include also comparatively subtle differences in the degree of segmental reduction, the intensity level, and the shape of intonation movements, especially of rises. Our study enhanced this production evidence and confirmed for Standard Northern German that listeners do in fact pick up on these additional phrase-final differences and interpret them – in parallel to their use in production – as cues to turn yielding and/or holding.

The question that we raised in 1.2 can thus be answered affirmatively; and this means that our study laid the ground for a broader scope in the phonetics of turn-taking. In particular, as is demonstrated by the turn-taking effects of segmental elision and assimilation, this broader scope has to span the traditionally separated segmental and prosodic layers of the speech signal. That is, like for prominence, intonation, and many other form-function systems, the phonetics of turn-taking is not merely a matter of prosody.

Moreover, our findings stress that understanding speech communication includes having a constant eye on phonetic detail. Every phonetic detail should initially be considered functional rather than prejudging it as epiphenomenal or random variation.

Previous studies, some of which are cited in 1.1, have shown that the production and perception of turn yielding and holding exhibit strong similarities across many – even unrelated – languages. For this reason, we assume that our findings are of general cross-linguistic significance. Testing this assumption will be the obvious next step. The corresponding perception experiments should use the same innovative task as the present study. Although this task is complex, its interactive concept proved to yield clear results while ensuring a high level of ecological validity.

References

E.L. Asu. 2006. Rising intonation in Estonian: an analysis of map task dialogues and spontaneous conversations. *Proc. Phonetic Symposium 2006, Helsinki, Finland*: 1-9.

W.G. Beattie. 1981. The regulation of speaker turns in face-to-face conversation: Some implications for conversation in sound-only communication channels. *Semiotica*, 34: 55-70.

M. Clayards, R.N. Aslin, M.K. Tanenhaus, and R.A. Jacobs. 2007. Within category phonetic variability affects perceptual uncertainty. *Proc. 16th International Congress of Phonetic Sciences, Saarbrücken, Germany:* 701-704.

C. Clemens and C. Diekhaus. 2009. Prosodic turn-yielding Cues with and without optical Feedback. *Proc. SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, London, UK*: 107–110.

G.J. Docherty, J. Milroy, L. Milroy, and D. Walshaw. 1997. Descriptive adequacy in phonology: A variationist perspective. *J. of Linguistics*, 33: 275‑310.

E. Dombrowski and O. Niebuhr. 2005. Acoustic patterns and communicative functions of phrase-final rises in German: activating and restricting contours. *Phonetica,* 62: 176-195.

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

J. Edlund and M. Heldner. 2005. Exploring prosody in interaction control. *Phonetica*, 62: 215-226.

J.-Y. Fon. 2002. *A cross-linguistic study on syntactic and discourse boundary cues in spontaneous speech*. PhD thesis, Ohio State University, USA.

J.-Y. Fon, K. Johnson, and S. Chen. 2011. Durational Patterning at Syntactic and Discourse Boundaries in Mandarin Spontaneous Speech. *Language and Speech*, 54: 5-32.

H. Friedberg. 2011. Turn-taking cues in a human tutoring cirpus. *Proc. Association for Computational Linguistics, Portland, USA*: 94-98.

A. Gravano. 2009. *Turn-taking and affirmative cue words in task-oriented dialogue*. PhD thesis, Columbia University, USA.

A. Gravano and J. Hirschberg. 2009. Turn yielding cues in task-oriented dialogue. *Proc. SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, *London, UK:* 253-261.

A. Kendon. 1995. Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics,* 23: 247-279.

K.J. Kohler. 1983. Prosodic boundary signals in German. *Phonetica*, 40: 89–134.

K.J. Kohler. 1996. Labelled data bank of spoken Standard German: The Kiel Corpus of Spontaneous Speech. *Proc. 4th International Conference on Spoken Language Processing, Philadelphia, USA*: 1938-1941.

K.J. Kohler. 2004. Categorical speech perception revisited. *Proc. of the Conference From Sound to Sense: 50+ years of discoveries in speech communication, MIT, Cambridge, USA*: 1-6.

K.J. Kohler and O. Niebuhr. 2011. On the role of articulatory prosodies in German message decoding. *Phonetica,* 68: 1-31.

D. H. Klatt. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America 59*: 1208-21.

I. Lehiste. 1982. Some phonetic characteristics of discourse. *Studia Linguistica*, 36: 117-130

J. Local, J. Kelly, and W.H. Wells. 1986. Towards a phonology of conversation: Turn-taking in Tyneside English. *Journal of Linguistics*, 22: 411–437.

J. Local and. G. Walker. 2012. How phonetic features project more talk. *Journal of the International Phonetic Association*, 42: 255-281.

C.H. Nakatani, J. Hirschberg, and B.J. Grosz. 1995. Discourse structure in spoken language: Studies on speech corpora. *Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, Palo Alto, USA*: 1-7.

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

B. Peters. 2006. *Form und Funktion prosodischer Grenzen im Gespräch*. PhD thesis, Christian-Albrechts-Universität zu Kiel, Germany.

A. Raux. 2008. *Flexible turn-taking for spoken dialog systems*. PhD thesis, Carnegie Mellon University, Pittsburgh, USA.

M. Taboada. 2006. Spontaneous and non-spontaneous turn taking. *Pragmatics*, 16: 329-360.

J. Vaissière, J. and A. Michaud. 2006. Prosodic constituents in French: a data-driven approach. In I. Fónagy, Y. Kawaguchi, T. Moriguchi (eds), *Prosody and syntax* (pp. 47-64). Amsterdam: John Benjamins.

K. Weilhammer and S. Rabold. 2003. Durational aspects in turn taking. *Proc. 15th International Congress of Phonetic Sciences, Barcelona, Spain* : 931–934.