Counting 'uhm's: How tracking the distribution of native and non-native disfluencies influences online language comprehension

Hans Rutger Bosker[a,b,*], Marjolein van Os[a,c], Rik Does[a,c], Geertje van Bergen[a,b]

[a] Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, the Netherlands
[b] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
[c] Radboud University, Nijmegen, the Netherlands

ABSTRACT

Disfluencies, like *uh*, have been shown to help listeners anticipate reference to low-frequency words. The associative account of this 'disfluency bias' proposes that listeners learn to associate disfluency with low-frequency referents based on prior exposure to non-arbitrary disfluency distributions (i.e., greater probability of low-frequency words after disfluencies). However, there is limited evidence for listeners actually tracking disfluency distributions online. The present experiments are the first to show that adult listeners, exposed to a typical or more atypical disfluency distribution (i.e., hearing a talker unexpectedly say *uh* before high-frequency words), flexibly adjust their predictive strategies to the disfluency distribution at hand (e.g., learn to predict high-frequency referents after disfluency). However, when listeners were presented with the same atypical disfluency distribution but produced by a non-native speaker, no adjustment was observed. This suggests pragmatic inferences can modulate distributional learning, revealing the flexibility of, and constraints on, distributional learning in incremental language comprehension.

Introduction

Humans are capable of rapidly detecting and adapting to statistical properties in their sensory environment. When it comes to language, this phenomenon, known as distributional or statistical learning, has been argued to underlie speech processing and language acquisition (Saffran, Aslin, & Newport, 1996). Emerging frameworks in psycholinguistics (e.g., P-chain model: Dell & Chang, 2013; Belief-updating model: Kleinschmidt & Jaeger, 2015) state that listeners overcome the variability and uncertainty that is so pervasive in spontaneous spoken communication by flexibly adjusting their expectations about future input based on prior experiences. Distributional learning within the scope of a single experimental session has been demonstrated in the online processing of prosodic (Kurumada, Brown, & Tanenhaus, 2012), phonological (Norris, McQueen, & Cutler, 2003), syntactic (Fine, Jaeger, Farmer, & Qian, 2013), and semantic cues (Yildirim, Degen, Tanenhaus, & Jaeger, 2016) in language. The current study contributes to these psycholinguistic perspectives by showing, for the first time, that listeners also track the distribution of metalinguistic performance cues, namely disfluencies like *uh*.

At the same time, listeners draw pragmatic inferences about the speaker and communicative situation at hand, which may influence how linguistic cues are weighed (Bosker & Reinisch, 2015; Brunellière & Soto-Faraco, 2013; Grodner & Sedivy, 2011; Kleinschmidt & Jaeger, 2015; Kraljic, Samuel, & Brennan, 2008; van Bergen & Bosker, 2018; Van Berkum, Van den Brink, Tesink, Kos, & Hagoort, 2008). For instance, while hearing a native talker make a grammatical error results in a P600 effect in listeners' ERPs, no such effect is observed when the same error is produced by a non-native talker (Hanulíková, Van Alphen, Van Goch, & Weber, 2012). The present experiments demonstrate that distributional learning in disfluency processing is modulated by pragmatic inferences about the talker, showing that listeners differentially adjust to the same disfluency distribution in native vs. non-native speech.

Spoken communication in everyday situations is riddled with disfluencies, including silent pauses, filled pauses or fillers (e.g., *uh*), repetitions, corrections, etc. It is estimated that about 6 disfluencies are produced every 100 words (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Fox Tree, 1995). Disfluencies follow a non-arbitrary distribution in native speech, with a greater likelihood of occurrence before relatively more complex lexical items, such as low-frequency words (De Jong, 2016; Hartsuiker & Notebaert, 2010; Levelt, 1983).

Listeners have been argued to be sensitive to the statistical regularities in the disfluency distribution in native speech, as evidenced by

---

prediction of more complex referents upon hearing the filler *uh*. For instance, in a seminal eye-tracking study by Arnold, Hudson Kam, and Tanenhaus (2007), listeners were presented with fluent ("Click on the red [target]") and disfluent instructions ("Click on thee uh… red [target]") to click on either known (e.g., an ice-cream cone) or unknown objects (unidentifiable symbols lacking a conventionalized lexical label). Recordings of participants' eye movements showed that listeners already anticipated reference to unknown objects (increase in proportion of fixations to unknown objects) when hearing the filler *uh* in the disfluent condition (i.e., well before hearing the target). Other studies have since shown the same effect in children as young as 2 years of age (Kidd, White, & Aslin, 2011; Orena & White, 2015; Owens & Graham, 2016; Owens, Thacker, & Graham, 2018; Thacker, Chambers, & Graham, 2018a, 2018b). Adult listeners have also been shown to be able to predict other types of complex referents, such as discourse-new (Arnold, Fagnano, & Tanenhaus, 2003; Arnold, Tanenhaus, Altmann, & Fagnano, 2004; Barr & Seyfeddinipur, 2010), compound (Watanabe, Hirose, Den, & Minematsu, 2008), and low-frequency referents (Bosker, Quené, Sanders, & De Jong, 2014a) upon hearing a disfluent filler *uh*.

Listeners' so-called *disfluency bias* for relatively more complex referents has been attributed (at least in part) to prior exposure to the non-arbitrary distribution of disfluencies in native speech (Bosker et al., 2014a). Given that disfluencies tend to occur more often before complex lexical items than before relatively simple words (Arnold & Tanenhaus, 2011; De Jong, 2016; Hartsuiker & Notebaert, 2010), listeners can cleverly use disfluencies as a predictive cue in comprehension. We refer to this account as the *associative account* of the disfluency bias: listeners anticipate more complex referents upon encountering a disfluency due to association with the non-arbitrary disfluency distribution in spontaneous speech.

To date, however, there is little evidence for distributional learning in the online processing of disfluencies. For instance, in most of the studies investigating the disfluency bias the factor Fluency was fully crossed with the particular target manipulation (e.g., known and unknown objects occurred equally often in fluent and disfluent utterances; Arnold et al., 2007). If listeners can flexibly adjust to this particular disfluency distribution over the course of an experiment, one would expect the disfluency bias for unknown objects to decrease as the experiment progresses (since, in the experiment, the disfluency is not predictive of the target to follow). However, none of the eye-tracking studies cited in this paper report any order effect (neither positive nor negative; cf. Corley & Hartsuiker, 2011 for an overall order effect in reaction time data, but no interaction between order and delay), except for Bosker et al. (2014a) and Thacker, Chambers, and Graham (2018b).

In Bosker et al. (2014a), participants were presented with fluent and disfluent instructions to click on either a high-frequency (e.g., a hand) or a low-frequency object (e.g., an igloo). Participants were reported to anticipate low-frequency objects upon hearing the filler *uh* in disfluent utterances. Moreover, this disfluency bias was found to *increase* over the course of the experiment, which the authors explained with reference to the disfluency distribution in their experiment. Even though fluency was fully crossed with word frequency in experimental trials, in filler trials participants consistently heard low-frequency words in disfluent utterances and high-frequency words in fluent utterances. As such, the overall disfluency distribution across the entire experiment was skewed towards low-frequency objects, which the authors speculated the participants may have adjusted to.

In Thacker et al. (2018b), children were presented with two talkers with distinct favorite colors (blue for male talker, pink for female talker). In their first experiment, the children were found to anticipate reference to objects of that talker's dispreferred color in disfluent utterances (and *vice versa* in fluent utterances). This disfluency bias, however, *diminished* slightly over the course of the experiment. To test whether this was a result of adjustment to the disfluency distribution in the experiment, they increased the percentage of 'disconfirming' trials (dispreferred colors in fluent utterances and preferred colors in

disfluent utterances) in a second experiment. Consequently, the disfluency bias towards dispreferred colors diminished even faster.

The manipulation of disfluency distributions in Thacker et al. (2018b) is, to date, the strongest evidence for distributional learning in the processing of disfluencies. However, its support for the associative account of the disfluency bias is limited by three factors. First, it showed a *reduction* in the disfluency bias across the experiment, without showing how a disfluency bias can gradually arise. Second, it showed modulation of a disfluency bias for talker color preferences that were established only moments before the experiment started. It may be argued that this disfluency bias is more susceptible to modulation, because associations are built on only a few examples. And third, it studied 5-year old children; as such, it remains unknown whether plasticity in disfluency processing is sustained in adults.

Given limited evidence for distributional learning in disfluency processing so far, some studies have actually argued against the associative account. Their arguments mostly hinged on flexibility in the disfluency bias. For instance, prior knowledge about talker identity can reduce the disfluency bias. When participants are told up front that they will listen to an atypical talker with 'object agnosia', who experiences trouble in naming even ordinary objects, the disfluency bias for unknown objects disappears (Arnold et al., 2007). Also, the disfluency bias for discourse-new referents only arises when the referents are new for the particular speaker at hand (Barr & Seyfeddinipur, 2010; but see Heller, Arnold, Klein, & Tanenhaus, 2014). These findings cannot easily be accounted for by the associative account and instead support an *inferential account,* where listeners infer the cause of production difficulty in the particular speaker to guide their visual attention during listening.

A study particularly relevant for the present investigation is Bosker et al. (2014a), who, as mentioned above, showed a disfluency bias for low-frequency referents when listening to native speech. A second experiment tested native listeners' processing of disfluencies produced by a non-native speaker, who produced the same fluent and disfluent instructions, but with a strong foreign accent. Listeners' gaze patterns revealed that there was no disfluency bias when listening to a non-native talker: listeners were as likely to look at high- or low-frequency objects when hearing a non-native *uh*. The authors explained the absence of a disfluency bias for non-native speech in terms of the atypical distribution of non-native disfluencies. Non-native speakers typically produce fewer low-frequency words, resorting to more commonly used lexical items (De Jong, 2016), and more disfluencies, such as silent and filled pauses (Bosker, Quené, Sanders, & De Jong, 2014b), negatively affecting non-native fluency perception (Bosker, Pinget, Quené, Sanders, & De Jong, 2013). As such, non-native disfluencies occur more often before high-frequency words than they do in native speech. Consequently, listeners, upon hearing a foreign accent in non-native speech, may bring this less regular non-native disfluency distribution to bear, realize that non-native disfluencies are less indicative of the word to follow, and thus show no disfluency bias when listening to non-native speech.

This explanation for the absence of a disfluency bias for non-native speech is in line with the associative account: in non-native speech, the association between disfluencies and low-frequency referents is simply much weaker. However, as explained above, evidence for adjustments in the processing of disfluencies based on prior exposure to particular disfluency distributions is limited. In fact, the same results may be explained by assuming that listeners abandon all predictions entirely when listening to non-native speech. This is supported by studies finding that other forms of prediction (phonological, syntactic, etc.) are also attenuated in non-native speech comprehension (Brunellière & Soto-Faraco, 2013; Hanulíková et al., 2012). Maybe listening to foreign accented speech is so cognitively taxing (Quené & Van Delft, 2010; Van Wijngaarden, 2001) and slow (Munro & Derwing, 1995) that it prevents listeners from using predictive strategies in disfluency processing altogether, regardless of disfluency distributions.
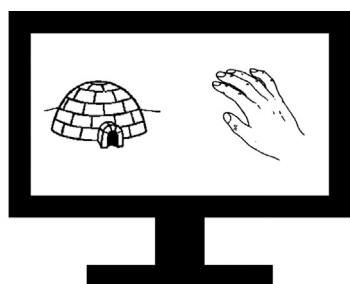
Fig. 1. Example display of an image pair, consisting of one low-frequency (igloo) and one high-frequency object (hand). Position on screen of the low-frequency and high-frequency item (left vs. right) was counter-balanced across trials. Participants in the Typical Group heard disfluent instructions to click on low-frequency objects (LF) and fluent instructions to click on high-frequency objects (HF). Conversely, participants in the Atypical Group heard disfluent instructions to click on high-frequency objects and fluent instructions to click on low-frequency objects.

The present study has two aims: first, it investigates whether listeners track the distribution of disfluencies in an online fashion during the course of a single experiment, flexibly adjusting their predictive strategies based on the disfluency distribution at hand. Experiment 1 – modelled after Bosker et al. (2014a) – involved eye-tracking (visual world paradigm) and was designed to address this first question. Two groups of Dutch participants were presented with fluent (*Klik op de [target]* "Click on the [target]") and disfluent instructions (*Klik op uh… de [target]* "Click on uh… the [target]"), produced by a native speaker of Dutch, to click on high-frequency (e.g., a hand) and low-frequency (e.g., an igloo) referents presented on screen, while their eye fixations were recorded (see Fig. 1). Crucially, the two groups differed in the disfluency distribution they were exposed to. The Typical Group consistently heard the native speaker produce high-frequency targets in fluent utterances and low-frequency targets in disfluent utterances, reflecting (to some degree) the typical distribution of disfluencies in native speech. The Atypical Group heard the same speaker produce the opposite disfluency distribution: high-frequency targets appeared in disfluent utterances and low-frequency targets in fluent utterances.

If listeners can flexibly adjust their predictive strategies in an online fashion, based on exposure to particular disfluency distributions over the course of an experimental session, we predict differential looking behavior for the two groups. The Typical Group will likely start out with some disfluency bias to low-frequency referents at the start of the experiment (based on life-long exposure to the non-arbitrary disfluency distribution in native Dutch), which would increase over the course of the experiment due to exposure to the disfluency distribution at hand. The Atypical Group may also start out with a disfluency bias toward low-frequency referents in early trials, but, as a result of more and more exposure to the atypical disfluency distribution, this may actually change to a disfluency bias toward *high-frequency* referents in later trials. Alternatively, if listeners are not sensitive to the distribution of disfluencies in native speech, no group differences would be expected.

As described in detail below, the procedure of Experiment 1 involved presenting participants with a set of image pairs, which was repeated four times across an experimental session. The advantage of this design was that 'early' trials (e.g., in block 1) were identical to 'later' trials (e.g., in block 4), facilitating comparison of the disfluency bias as the experiment progressed. However, repetitions of items may have allowed for item-level priming, which could possibly serve as an alternative explanation for the findings in Experiment 1. Therefore, Experiment 2 was identical to Experiment 1 without repeating items in

an experimental session. Thus, Experiment 2 assessed the contribution of item-level priming to the results of Experiment 1.

The second aim of the present study is to investigate whether adaptation in the processing of disfluencies, induced by distributional learning, is similar for native and non-native speech. Recall that Bosker et al. (2014a) did not observe a disfluency bias for low-frequency referents when listening to a non-native speaker, which they explained in terms of the more irregular disfluency distribution in non-native speech. This interpretation predicts that, when listening to a non-native speaker with a highly regular disfluency distribution, we may expect a disfluency bias in non-native speech to arise after all. Alternatively, if the absence of a disfluency bias is due to attenuated prediction in non-native speech in general, we should not find a disfluency bias for any non-native speaker.

Experiment 3 was designed to address this second question. It was identical to Experiment 1 (two groups listening to fluent and disfluent instructions) except that this time all materials were produced by a non-native speaker of Dutch with a strong foreign accent. If listeners adjust their predictive strategies based on the particular disfluency distributions at hand, regardless of talker identity, we should find similar learning behavior in Experiment 3 as in Experiment 1 (i.e., gradually stronger disfluency bias for low-frequency referents in Typical Group; gradually opposite disfluency bias in Atypical Group). If, however, non-native disfluencies do not induce any prediction, neither groups in Experiment 3 should show any disfluency bias.

Experiment 1

Method

Participants. Native Dutch participants ($N = 41$) with normal hearing were recruited from the Max Planck Institute's participant pool. Participants in all experiments reported in this study gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW2014-1003-196). Data from eleven participants were excluded due to trouble with eye-tracking calibration ($n = 8$) and due to technical issues with audio presentation ($n = 3$), leaving a total of 30 participants for analysis (23 females, 7 males; mean age = 22; range = 19–33). Half of these participants were assigned to the Typical Group, the other half to the Atypical Group (see *procedure*).

Materials and design. Materials were adapted from Bosker et al. (2014a). These involved a set of 30 low-frequency (LF; mean log frequency = 0.38, $SD = 0.28$) and 30 high-frequency objects in the form of line drawings (HF; mean log frequency = 2.07, $SD = 0.29$) taken from Severens, Lommel, Ratinckx, and Hartsuiker (2005). All images had high name agreement (LF = 96.7, $SD = 3.64$; HF = 97.3, $SD = 3.49$), were of common gender selecting the Dutch article *de*, and were scaled to have a maximal length/width of 300 pixels. Each LF image was paired four times with a different HF image, avoiding phonological overlap between pair members, resulting in 4 unique sets of 30 LF-HF image pairs. For a complete list, see Table S1 in the Supplementary Material.

In Bosker et al. (2014a), a female native speaker of Dutch had been recorded, producing each image label in a fluent sentence frame (*Klik op de [target]* "Click on the [target]") and in a disfluent sentence frame (*Klik op uh… de [target]* "Click on uh… the [target]"). From these recordings, three natural sounding fluent sentence tokens (*Klik op*) and three disfluent sentence tokens (*Klik op uh…*) were excised. Also, target fragments including the article *de* were excised (i.e., all speech from the onset of the article *de* up to the offset of the target label), half from a fluent recording, half from a disfluent recording. These target fragments were randomly assigned to one of the three fluent sentence tokens (resulting in a fluent instruction), and to one of the three disfluent sentence tokens (resulting in a disfluent instruction). Note that, as a consequence of this cross-splicing, target words were identical across

fluent and disfluent conditions. Also note that the difference between the fluent and disfluent condition was located in the sentence frame, with the disfluent condition additionally containing the filler *uh*, as well as differing in several other prosodic characteristics (e.g., longer syllable durations, higher pitch; see Bosker et al., 2014a).

Procedure. Participants were tested individually in a sound-attenuating booth. They were seated at a distance of approximately 60 cm in front of a screen with a remote EyeLink 1000 eye-tracking system (SR Research) and listened to stimuli over headphones at a comfortable volume. Stimulus presentation was controlled by Experiment Builder software (SR Research), sampling the right eye at 500 Hz.

Each session started with telling the participant a cover story about the purpose of the eye-tracking experiment. The experiment supposedly tested the extent to which listeners could correctly follow up instructions from various speakers. Purportedly, recordings had been made of 20 speakers, including both native and non-native speakers of Dutch, who had been instructed to name images using a standard instruction template, namely *Klik op de [target]* "Click on the [target]". This cover story was motivated by the need to justify the presence of disfluencies in the speech. Moreover, it meant that listeners might plausibly attribute the disfluencies to difficulty in word retrieval. Participants in Experiment 1 were explicitly told they would be listening to speech from a native speaker.

Half of the participants had been assigned to the Typical Group, the other half to the Atypical Group. The experimental procedure was identical for both groups: all participants were presented with 50% fluent (*Klik op…*) and 50% disfluent spoken instructions (*Klik op uh…*) to click on one of two images. Crucially, the two groups differed in the distribution of LF and HF targets across fluent and disfluent conditions. The Typical Group exclusively heard fluent instructions followed by HF targets and disfluent instructions followed by LF targets. As such, the presence of the filler *uh* was a highly reliable cue to an upcoming LF referent. Conversely, the Atypical Group exclusively heard fluent instructions combined with LF targets and disfluent instructions combined with HF targets. Hence, in this group, the filler *uh* was a highly reliable cue to an upcoming HF referent.

Before starting the eye-tracking experiment, participants were familiarized with the images and their labels. All images used in the main task were shown to the participants together with the label used to refer to the images. After 3–6 trials (randomly selected), participants' attention and recognition accuracy was verified by presenting them one image from the previous 3–6 trials and asking them to type in the corresponding label.

The eye-tracking experiment started with a nine-point calibration procedure. On each trial, participants first had to click with the computer mouse on a blue rectangle in the middle of the screen to center their eye gaze and mouse position. This screen was immediately followed by two images, centered in the left and right halves of the screen (see Fig. 1). The position of LF and HF images (left or right) was randomized and counter-balanced across participants. After a 2 s preview, the auditory stimulus was presented. Participants were instructed to listen to and follow up the spoken instructions using a standard computer mouse. The images stayed on the screen until the participant responded by clicking on one of the presented images. After an intertrial interval of 1 s, the next trial started automatically.

Participants were presented with the four sets of 30 image pairs in sequence (total of 120 trials), with a random order of trials within each block. Following the eye-tracking experiment, participants were debriefed about what they thought the purpose of the experiment was, whether they had noticed a correlation between disfluency and target frequency, and the extent to which they believed the cover story. Although many participants noticed the relatively high disfluency incidence, none of the participants reported awareness of the disfluency distribution they had been allocated to.
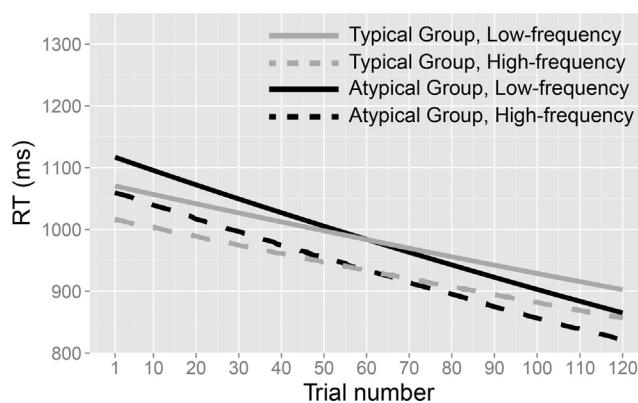

Fig. 2. Modelled reaction times (RTs) in ms from Experiment 1 (native speaker) as a function of the trial number, split by target frequency (solid lines show low-frequency targets, dashed lines show high-frequency targets) and group (Typical Group in gray, Atypical Group in black), as predicted by the Linear Mixed Effects Regression model.

Results

Mouse clicks. Participants were very accurate in their mouse clicks: only in 7 trials ($< 0.2\%$) did participants select the competitor image over the target image. Before analyzing the reaction times (RTs), trials with incorrect clicks or outlier RTs ($> 2SD$ above the mean; $n = 56$; $< 2\%$) were excluded from analyses. Raw reaction times (RTs), calculated from target word onset (in ms), from Experiment 1 are given in Fig. S1 in the Supplementary Material. Mean (SD) reaction times were: $HF_{Typical} = 940$ (318); $LF_{Typical} = 989$ (322); $HF_{Atypical} = 965$ (278); $LF_{Atypical} = 1013$ (291); see also Fig. 2.

Log-transformed RTs were analyzed using a Linear Mixed Effects Regression analysis (Baayen, Davidson, & Bates, 2008) as implemented in lme4 library (version 1.0.5; Bates, Maechler, Bolker, & Walker, 2015) in R (R Development Core Team, 2012). The fixed effects structure included Group (categorical predictor with two levels using deviation coding: Typical coded as -0.5, Atypical as $+0.5$), Frequency (categorical predictor with two levels using deviation coding: High Frequency coded as -0.5, Low Frequency as $+0.5$), Trial Number (continuous predictor, z-scored to improve convergence), and all interactions. By-participant and by-item random intercepts were included; adding random slopes for fixed effects resulted in non-convergence of models. Statistical significance was assessed at the 0.05 significance level by checking whether effects had absolute *t*-values exceeding 2 (Baayen, 2008).

This model revealed a main effect of Frequency ($\beta = 0.052$, $SE = 0.020$, $t = 2.616$), indicating longer RTs for LF targets; and a main effect of Trial Number ($\beta = -0.050$, $SE = 0.005$, $t = -10.323$), indicating shorter RTs for later trials. No overall difference between the two groups was found, but an interaction between Group and Trial Number ($\beta = -0.025$, $SE = 0.009$, $t = -2.633$) revealed a more pronounced order effect for the Atypical Group. This suggests that the Atypical Group started out with slower responses than the Typical Group, but RTs decreased more in the Atypical Group than in the Typical Group as the experiment progressed (cf. Fig. 2), likely as a result of adaptation to the atypical disfluency distribution.

Eye fixations – Disfluent speech. Prior to the analyses, blinks and saccades were excluded from the data. Only fixations on the images themselves were coded as looks toward a particular image. The eye fixation data were downsampled to 250 Hz for simplicity.

The fluent and disfluent condition differed considerably in the time at which certain words were presented, which prevented us from combining the data from the fluent and disfluent conditions into one analysis. Therefore, separate analyses were run per fluency condition (cf. Bosker et al., 2014a). Also, as this study focused on *anticipation* of
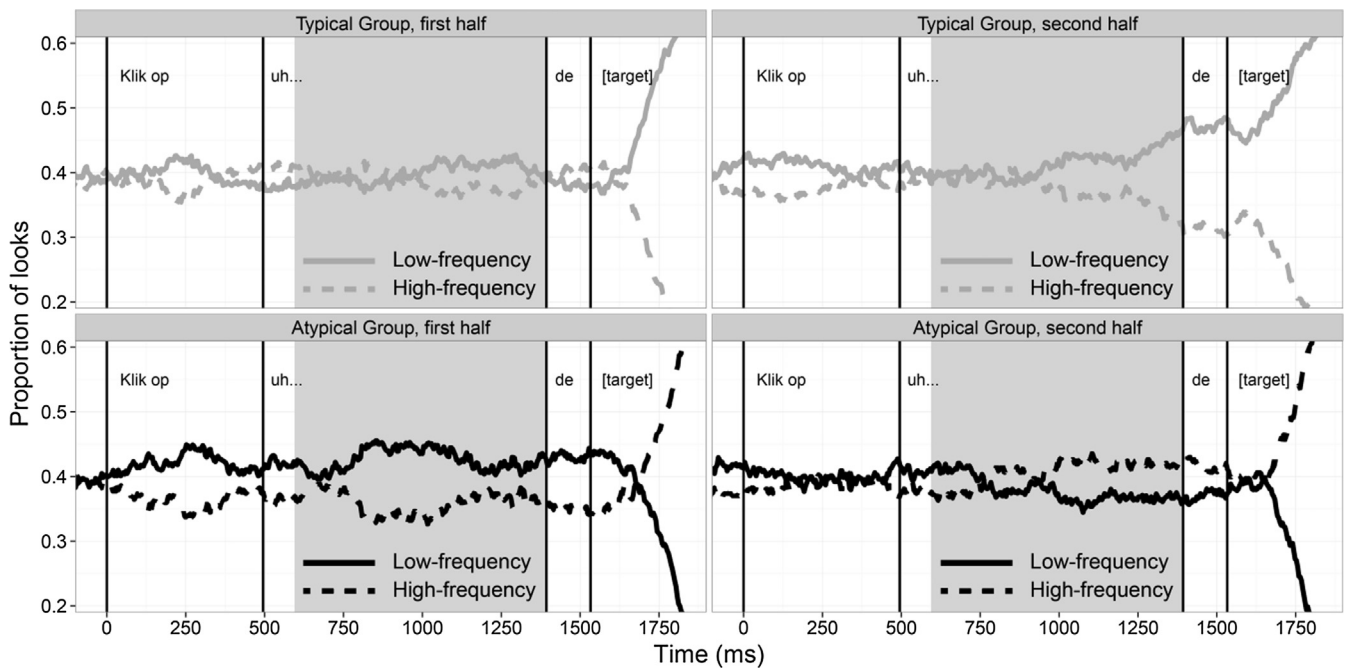
Fig. 3. Proportion of looks, in disfluent trials only, from Experiment 1 (native speaker) to high-frequency objects (dashed lines) and low-frequency objects (solid lines), separately for the two groups and the two halves of the experiment. Time in ms is calculated from sound onset. Vertical lines represent the (mean) onsets of words in the sentence. The gray rectangle indicates the time window of interest as identified for the statistical analyses. Note that the data are shown separately for the first and second half for illustration purposes only; all statistical analyses involved the predictor Trial Number.

linguistic content as evidenced in eye fixations, gaze patterns were evaluated in time windows before any acoustic information regarding the target had been encountered.

In the disfluent condition, the time window of interest started from *uh* onset + 100 ms (to correct for the minimal time it takes to launch a saccade; Altmann, 2011) up to article onset (not target onset, because the article *de* may already have contained coarticulatory cues to the target referent); see Fig. 3 (mean length: 798 ms). Note that Fig. 3 shows the data separately for the first and second half for illustration purposes only, to allow visual inspection of distributional learning for the two groups; all statistical analyses involved Trial Number as a continuous predictor.

We statistically tested the binomial variable fixations on low-frequency referents (0 = no, 1 = yes) in the disfluent condition at a sample-by-sample level using a Generalized Linear Mixed Model (GLMM; Quené & Van den Bergh, 2008) with a logistic linking function as implemented in the lme4 library in R. We selected this analysis method because it allows for crossed random effects for participants and items (Quené & Van den Bergh, 2008), it can model both the time course on a given trial as well as order effects across trials surfacing as the interaction between Time and Trial Number (Mirman, Dixon, & Magnuson, 2008), and because similar analysis methods were used in Bosker et al. (2014a). This particular GLMM included fixed effects of Group (categorical predictor with two levels using dummy coding, mapping the Typical Group onto the intercept), Time (continuous predictor, z-scored to improve convergence), Trial Number (continuous predictor, z-scored to improve convergence), and all interactions. By-participant and by-item random intercepts were included, but no random slopes since this resulted in non-convergence of models.

Note that the Typical Group was mapped onto the intercept and, as such, simple effects should be interpreted with respect to the Typical Group only. The model revealed significant effects of Time ($\beta = 0.054$, $SE = 0.005$, $z = 10.264$, $p < 0.001$; the proportion of looks to low-frequency referents gradually increased in the time window of interest for the Typical Group), no effect of Trial Number ($p = 0.292$), but a significant interaction between Time and Trial Number ($\beta = 0.046$,

$SE = 0.005$, $z = 8.627$, $p < 0.001$; a steeper increase in proportion of looks to low-frequency referents in the time window of interest as the experiment progressed).

Interactions with the predictor Group revealed differential effects in the Atypical Group. An interaction between Group and Time ($\beta = -0.080$, $SE = 0.007$, $z = -10.865$, $p < 0.001$) showed a relative *decrease* in the proportion of looks to low-frequency referents in the time window of interest in the Atypical Group. An interaction between Group and Trial Number ($\beta = -0.083$, $SE = 0.007$, $z = -11.179$, $p < 0.001$) showed that, in contrast to the Typical Group, the overall proportion of looks to low-frequency referents decreased as the experiment progressed in the Atypical Group. Finally, a three-way interaction between Group, Trial Number, and Time ($\beta = -0.084$, $SE = 0.007$, $z = -11.408$, $p < 0.001$) showed a steeper decrease in proportion of looks to low-frequency referents in the time window of interest as the experiment progressed in the Atypical Group.

Finally, Fig. 3 seems to show some unexpected gaze patterns prior to *uh* onset in the two groups. For instance, the Atypical Group would seem to have a preference for the LF referent in this early time window in the first half of the experiment, while the Typical Group seems to have a slight preference in this early time window in the second half of the experiment. To statistically assess potential early looking preferences, we calculated logit transformed fixation proportions to the LF and HF referents in the early time window from sound onset up to *uh* onset for each participant. These were tested statistically using a linear mixed model with the predictors Trial Number, Group, and Frequency (categorical predictor, low vs. high), and all their interactions, with by-participant random intercepts. No effects were found for any of the predictors nor any interactions (all $t < 1.2$), suggesting that these early patterns in Fig. 3 were spurious.

Eye fixations – Fluent speech. In the fluent condition, the time window of interest started from sound onset + 100 ms (correcting for saccade launch time) up to article onset, because, as mentioned above, the article may have contained coarticulatory cues to the target referent (see Fig. 4; mean length: 267 ms). Note that this involved a considerably shorter time window compared to the disfluent condition, as a
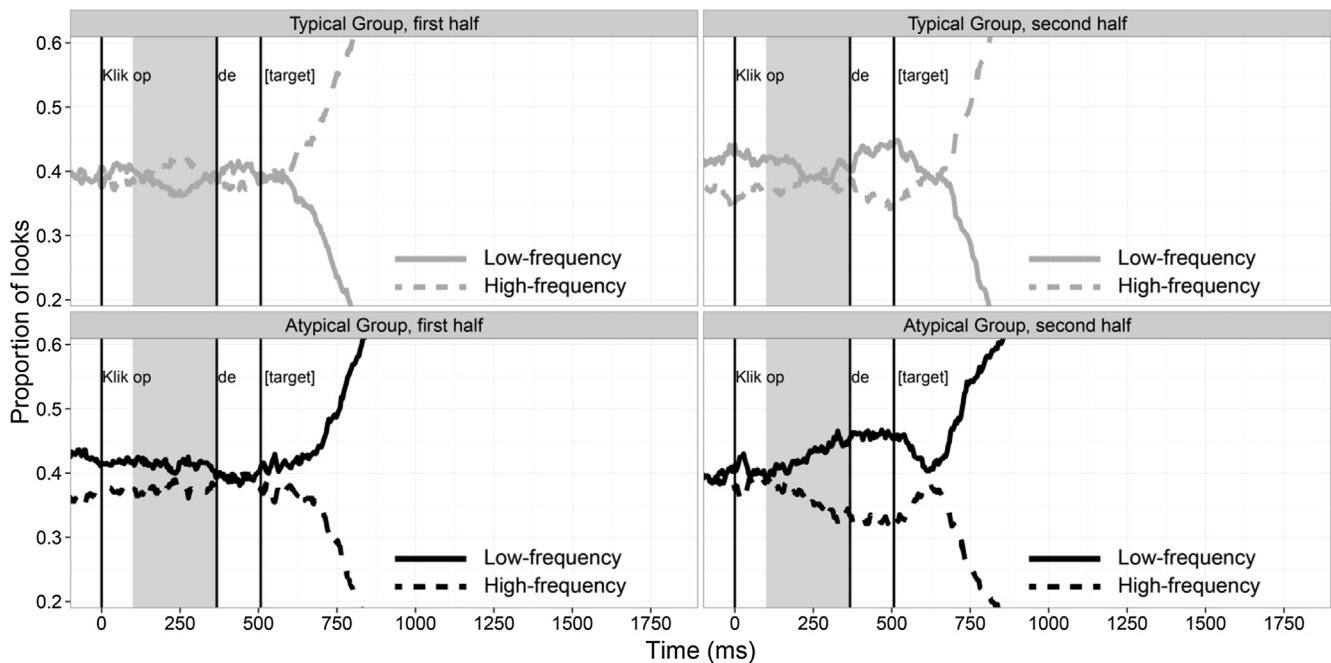
Fig. 4. Proportion of looks, in fluent trials only, from Experiment 1 (native speaker) to high-frequency objects (dashed lines) and low-frequency objects (solid lines), separately for the two groups and the two halves of the experiment. Time in ms is calculated from sound onset. Vertical lines represent the (mean) onsets of words in the sentence. The gray rectangle indicates the time window of interest as identified for the statistical analyses. Note that the data are shown separately for the first and second half for illustration purposes only; all statistical analyses involved the predictor Trial Number.

consequence of the more fluent speaking style.

Since we hypothesized that disfluent speech might induce anticipatory looks to low-frequency referents in the Typical Group, we conversely predict that fluent speech might induce anticipatory looks to high-frequency referents in the Typical Group (and *vice versa* for the Atypical Group). We statistically tested participants' fixations to *high-frequency* referents in this time window. The binomial variable fixations on high-frequency referents (0 = no, 1 = yes) in the fluent condition at a sample-by-sample level were entered into a GLMM with a logistic linking function as implemented in the lme4 library in R. This GLMM followed an identical structure as the GLMM for the disfluent condition, including the predictors Group, Time, Trial Number, and all interactions.

Note that the Typical Group was mapped onto the intercept; simple effects should thus be interpreted with respect to the Typical Group only. The model revealed an effect of Time ($\beta = 0.022$, $SE = 0.009$, $z = 2.422$, $p = 0.015$), indicating that the proportion of looks to high-frequency referents gradually increased over the course of the time window for the Typical Group. In addition, we found a negative effect of Trial Number ($\beta = -0.029$, $SE = 0.009$, $z = -3.108$, $p = 0.002$), showing that the overall proportion of looks to high-frequency referents gradually decreased over the course of the experiment for the Typical Group. We found no evidence for an interaction between Time and Trial Number ($p = 0.613$), suggesting that the gaze patterns of the Typical Group in the time window of interest did not change for fluent trials over the course of the experiment.

Interactions with the predictor Group again revealed differential effects in the Atypical Group. An interaction between Group and Time ($\beta = -0.053$, $SE = 0.013$, $z = -4.059$, $p < 0.001$) showed that the proportion of looks to high-frequency referents *decreased* across the time window of interest in the Atypical Group. An interaction between Group and Trial Number ($\beta = 0.046$, $SE = 0.013$, $z = 3.565$, $p < 0.001$) showed that, unexpectedly, the overall proportion of looks to high-frequency referents *increased* as the experiment progressed in the Atypical Group. However, a three-way interaction between Group, Trial Number, and Time ($\beta = -0.034$, $SE = 0.013$, $z = -2.613$,

$p = 0.009$) showed a steeper decrease in fixations to high-frequency referents over the course of the time window of interest as the experiment progressed in the Atypical Group.

Discussion

The data from Experiment 1 demonstrated differential distributional learning in the Typical vs. the Atypical Group. Upon encountering a disfluency, the Typical Group showed a slight preference to fixate low-frequency referents at the outset of the experiment, which considerably increased as the experiment progressed. This suggests that the Typical Group used the more typical disfluency distribution to adjust their predictive strategies during the experiment. Note, however, that we did not find evidence for anticipation of *high-frequency* referents in fluent trials for the Typical Group, but the time window allowing for anticipatory looking behavior was of course much shorter. Furthermore, the fluent trials could be argued to be 'less salient' than the disfluent trials, because the latter contained an explicit filled pause *uh*. Adaptation to 'implicit' fluency may be argued to be more difficult than adaptation to 'explicit' disfluency.

The Atypical Group started out with the same preference for low-frequency referents in disfluent trials, likely reflecting the typical disfluency distribution in natural speech. However, the Atypical Group adjusted their predictive strategies in the opposite direction: as a consequence of the atypical disfluency distribution in the experiment, the Atypical Group learnt to anticipate *high-frequency* referents upon hearing a disfluency in later trials. This distributional learning was also reflected in the fluent condition: there, the Atypical Group learnt to look away from the high-frequency referent (i.e., anticipate the low-frequency referent). Finally, additional support comes from the RT analysis. This showed that the Atypical Group started out with slower RTs compared to the Typical Group, presumably because of the atypical distribution of disfluencies. As the experiment progressed, the Atypical Group demonstrated a steeper decrease in response latencies compared with the Typical Group, suggesting stronger adaptation to the disfluency distribution at hand.

An anonymous reviewer suggested that the observed learning patterns in the Typical and Atypical Group in Experiment 1 may alternatively be explained in terms of item-level priming. Recall that a given image occurred as the target image in the spoken instructions four times in an experimental session (each time paired with a different competitor image). Hence, the results from Experiment 1 may also be argued to have arisen from item-level priming: if a participant heard "igloo" in a disfluent sentence in the beginning of the experiment, then hearing a disfluency later in the experiment could have primed the specific item "igloo". In order to investigate the contribution of item-level priming to the results of Experiment 1, Experiment 2 was identical to Experiment 1 except that no items were repeated in Experiment 2.

Experiment 2

Method

Participants. Native Dutch participants ($N = 40$) with normal hearing were recruited from the Max Planck Institute's participant pool. At debriefing, 3 participants reported to have detected the particular disfluency distribution they were exposed to in the experiment (2 from the Typical Group; 1 from the Atypical Group). Their data were excluded, leaving a total of 37 participants for analysis (29 females, 8 males; mean age = 22; range = 18–31). A number of 18 participants were assigned to the Typical Group, and 19 to the Atypical Group.

Materials and design. In Experiment 2, the image materials from Experiment 1 were re-used together with an additional set of images (more line drawings selected from Severens et al., 2005), such that repetitions of items were avoided. The complete set of 120 low-frequency (LF; mean log frequency = 0.19, SD = 0.25) and 120 high-frequency objects (HF; mean log frequency = 1.91, SD = 0.42) is given in Table S1 in the Supplementary Material. Each LF image was paired with a single HF image, avoiding phonological overlap between pair members, resulting in 120 unique LF-HF image pairs.

A new female native speaker of Dutch was recorded, producing each image label in fluent (*Klik op de [target]* "Click on the [target]") and disfluent sentence frames (*Klik op uh… de [target]* "Click on uh… the [target]"). These new native recordings were manipulated in Praat using the same procedure as in Experiment 1: article + target fragments were excised and cross-spliced onto six different (fluent and disfluent) sentence tokens.

Procedure. Each of the 120 unique LF-HF pairs was presented only once in a random order. The position of LF and HF images (left or right) was randomized and counter-balanced across trials and participants, as well as which of the two pair members was the target vs. competitor. The rest of the procedure of Experiment 2 was identical to Experiment 1, including the cover story, explicit instructions about listening to a native speaker of Dutch, the group design, and post-experimental debriefing. Several participants noticed the frequent occurrence of disfluencies in the experiment, but only 3 out of 40 participants explicitly reported (after the experiment) the (typical/atypical) disfluency distribution they had been exposed to. These participants were excluded from analyses.

Results

Mouse clicks. Participants were very accurate in their mouse clicks: only in 1 trial ($< 0.1\%$) did a participant select the competitor image over the target image. Before analyzing the reaction times (RTs), this incorrect trial and trials with outlier RTs ($> 2SD$ above the mean; $n = 113$; $< 3\%$) were excluded from analyses. Raw reaction times (RTs), calculated from target word onset (in ms), from Experiment 2 are given in Fig. S2 in the Supplementary Material. Mean (SD) reaction times were: $HF_{Typical}$ = 945 (208); $LF_{Typical}$ = 1038 (224); $HF_{Atypical}$ = 988 (269); $LF_{Atypical}$ = 1085 (277); see also Fig. 5.

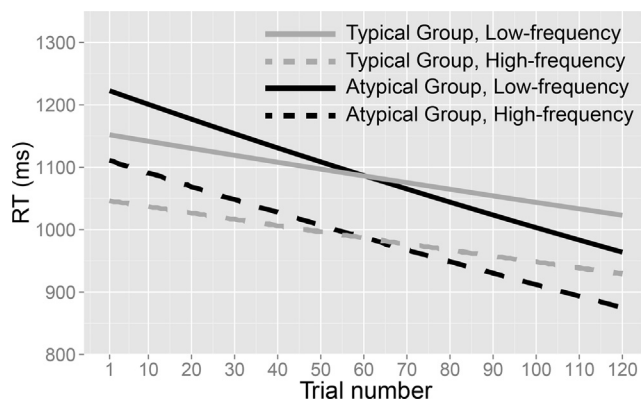A Linear Mixed Effects Regression analysis tested the log-

Fig. 5. Modelled reaction times (RTs) in ms from Experiment 2 (native speaker; no item repetitions) as a function of the trial number, split by target frequency (solid lines show low-frequency targets, dashed lines show high-frequency targets) and group (Typical Group in gray, Atypical Group in black), as predicted by the Linear Mixed Effects Regression model.

transformed RTs. This analysis involved the same structure as the one reported for Experiment 1, testing for effects of Group, Frequency, Trial Number, and all interactions. Statistical significance was assessed at the 0.05 significance level by checking whether effects had absolute t-values exceeding 2.

The results from this model were very similar to those in Experiment 1. The model revealed a main effect of Frequency ($\beta = 0.096$, $SE = 0.010$, $t = 9.500$), indicating longer RTs for LF targets; and a main effect of Trial Number ($\beta = -0.034$, $SE = 0.003$, $t = -11.710$), indicating shorter RTs for later trials. No overall difference between the two groups was found, but an interaction between Group and Trial Number ($\beta = -0.035$, $SE = 0.006$, $t = -5.890$) revealed a more pronounced order effect for the Atypical Group. This suggests that the Atypical Group started out with slower responses than the Typical Group, but RTs decreased more in the Atypical Group than in the Typical Group as the experiment progressed (cf. Fig. 5), likely as a result of adaptation to the atypical disfluency distribution.

Eye fixations – Disfluent speech. Prior to the analyses, blinks and saccades were excluded from the data. Only fixations on the images themselves were coded as looks toward a particular image. The eye fixation data were downsampled to 250 Hz for simplicity.

Similar to Experiment 1, the fluent and disfluent condition were analyzed separately. In the disfluent condition, the same time window as previously was used, starting from *uh* onset + 100 ms up to article onset; see Fig. 6 (mean length: 806 ms).

We statistically tested the binomial fixations on low-frequency referents in the disfluent condition using a GLMM with identical structure as the one reported in Experiment 1. Note that this also entailed that the Typical Group was mapped onto the intercept, meaning that main effects should be interpreted with respect to the Typical Group only.

This GLMM did not reveal a significant effect of Time ($\beta = -0.008$, $SE = 0.005$, $z = -1.461$, $p < 0.144$), suggesting that, overall, the proportion of looks to low-frequency referents did not gradually change in the time window of interest for the Typical Group. However, we did find an effect of Trial Number ($\beta = 0.043$, $SE = 0.005$, $z = 8.158$, $p < 0.001$), indicating an increase in the proportion of looks to low-frequency referents as the experiment progressed. Moreover, an interaction between Time and Trial Number ($\beta = 0.025$, $SE = 0.005$, $z = 4.869$, $p < 0.001$) demonstrated a steeper increase in proportion of looks to low-frequency referents in the time window of interest as the experiment progressed.

Interactions with the predictor Group revealed differential effects in the Atypical Group. An interaction between Group and Time ($\beta = -0.028$, $SE = 0.007$, $z = -3.861$, $p < 0.001$) showed a relative *decrease* in the proportion of looks to low-frequency referents in the
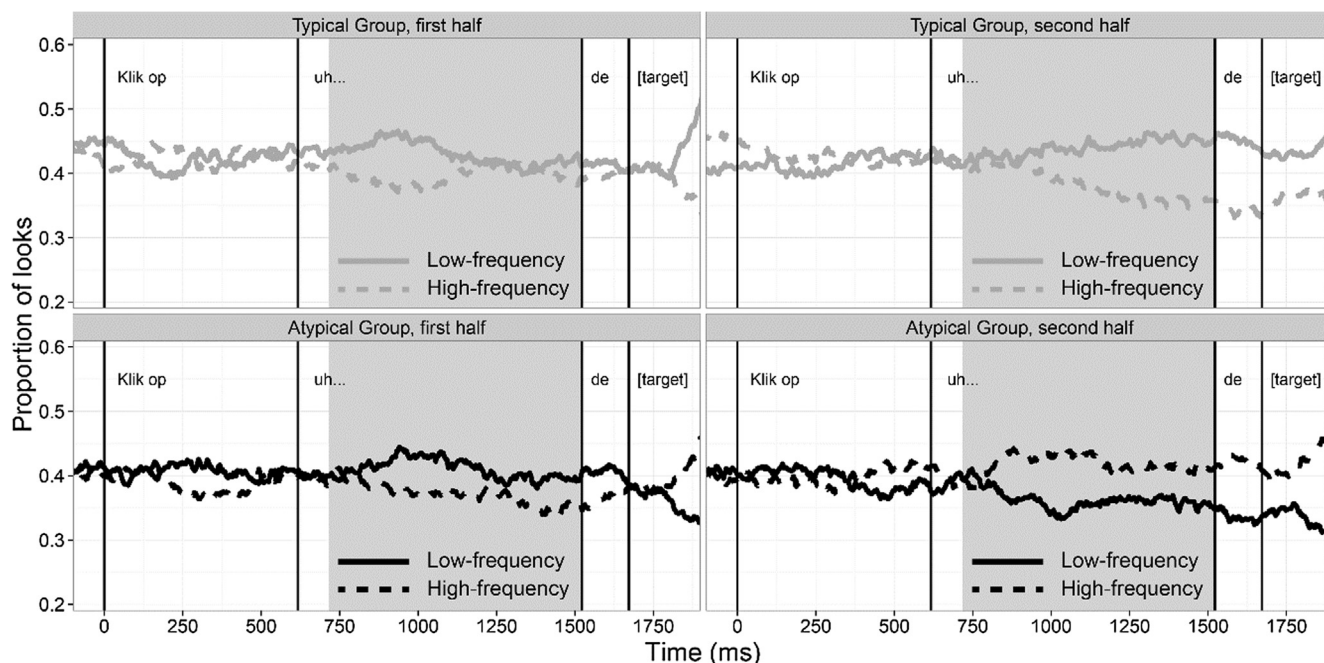
Fig. 6. Proportion of looks, in disfluent trials only, from Experiment 2 (native speaker; no item repetitions) to high-frequency objects (dashed lines) and low-frequency objects (solid lines), separately for the two groups and the two halves of the experiment. Time in ms is calculated from sound onset. Vertical lines represent the (mean) onsets of words in the sentence. The gray rectangle indicates the time window of interest as identified for the statistical analyses. Note that the data are shown separately for the first and second half for illustration purposes only; all statistical analyses involved the predictor Trial Number.

time window of interest in the Atypical Group. An interaction between Group and Trial Number ($\beta = -0.114$, $SE = 0.008$, $z = -15.068$, $p < 0.001$) showed that, in contrast to the Typical Group, the overall proportion of looks to low-frequency referents *decreased* as the experiment progressed in the Atypical Group. Finally, a three-way interaction between Group, Trial Number, and Time ($\beta = -0.016$, $SE = 0.007$, $z = -2.220$, $p = 0.026$) showed a steeper *decrease* in proportion of looks to low-frequency referents in the time window of interest as the experiment progressed in the Atypical Group.

Finally, similar to Experiment 1, we statistically tested for early looking preferences in the time window from sound onset up to *uh* onset for each participant. Again, no statistical evidence was found for early looking preferences for either HF or LF referents in either group.

Eye fixations – Fluent speech. In the fluent condition, the time window of interest started from sound onset + 100 ms up to article onset; see Fig. 7 (mean length: 422 ms).

We statistically tested participants' fixations on *high-frequency* referents in this time window by means of another GLMM with identical structure as used in Experiment 1. Note that the Typical Group was again mapped onto the intercept; simple effects should hence be interpreted with respect to the Typical Group only.

The model revealed an effect of Time ($\beta = 0.037$, $SE = 0.007$, $z = 4.973$, $p < 0.001$), indicating that the proportion of looks to high-frequency referents gradually increased over the course of the time window for the Typical Group. In addition, we found a negative effect of Trial Number ($\beta = -0.053$, $SE = 0.008$, $z = -6.626$, $p < 0.001$), showing that the overall proportion of looks to high-frequency referents gradually decreased over the course of the experiment for the Typical Group. Contrary to predictions, we also found a negative interaction between Time and Trial Number ($\beta = -0.038$, $SE = 0.007$, $z = -5.063$, $p < 0.001$), suggesting that the proportions of looks to high-frequency referents stabilized in the time window of interest for the Typical Group over the course of the experiment.

Interactions with the predictor Group revealed differential effects in the Atypical Group. An interaction between Group and Trial Number ($\beta = 0.051$, $SE = 0.011$, $z = 4.596$, $p < 0.001$) showed that the overall

proportion of looks to high-frequency referents *decreased* as the experiment progressed in the Atypical Group. No interaction between Group and Time ($p = 0.315$) and no three-way interaction between Group, Trial Number, and Time ($p = 0.958$) were observed.

Discussion

The results from Experiment 2 mirror those from Experiment 1. As the experiment progressed, participants in the Typical Group – upon encountering a disfluency – showed an increase in their preference to fixate low-frequency referents. This suggests, once again, that the Typical Group used the more typical disfluency distribution to adjust their predictive strategies during the experiment. In contrast, the Atypical Group adjusted their predictive strategies in the opposite direction: as a consequence of the atypical disfluency distribution, the Atypical Group learnt to anticipate *high-frequency* referents upon hearing a disfluency. Additional support comes from the RT data showing that the Atypical Group started out with slower RTs compared to the Typical Group, presumably because of the atypical distribution of disfluencies. As the experiment progressed, the Atypical Group demonstrated a steeper decrease in response latencies compared with the Typical Group, suggesting stronger adaptation to the disfluency distribution at hand. Crucially, these findings were obtained without item repetitions in Experiment 2, removing the possibility for item-level priming. Therefore, they show that the very similar outcomes of Experiment 1 are most likely explained in terms of distributional learning rather than item-level priming.

Experiment 3

Experiment 3 was designed to address the second question of the present study, investigating whether adaptation in the processing of disfluencies, induced by distributional learning, is similar for native and non-native speech. Note that Bosker et al. (2014a) did not find a disfluency bias for low-frequency referents when participants listened to a non-native speaker. Experiment 3 assessed whether a disfluency bias in
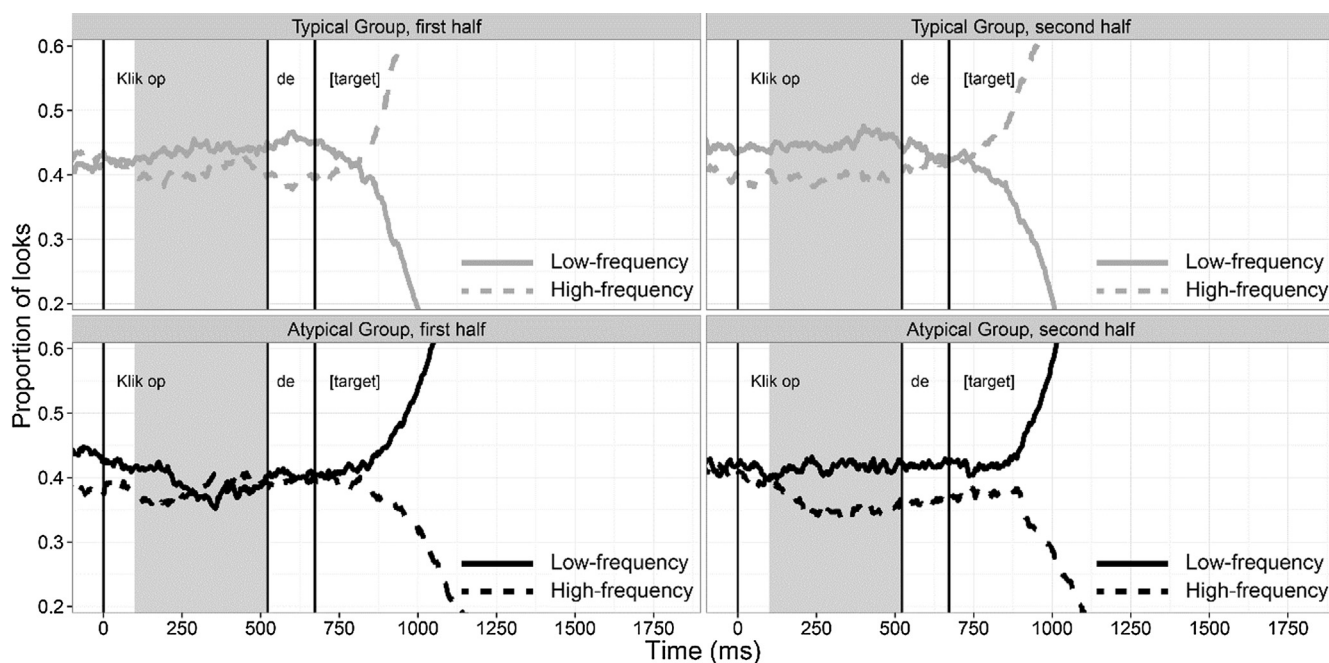
Fig. 7. Proportion of looks, in fluent trials only, from Experiment 2 (native speaker; no item repetitions) to high-frequency objects (dashed lines) and low-frequency objects (solid lines), separately for the two groups and the two halves of the experiment. Time in ms is calculated from sound onset. Vertical lines represent the (mean) onsets of words in the sentence. The gray rectangle indicates the time window of interest as identified for the statistical analyses. Note that the data are shown separately for the first and second half for illustration purposes only; all statistical analyses involved the predictor Trial Number.

non-native speech may arise after all if the non-native speaker at hand has a highly regular disfluency distribution. In order to increase our chances of finding evidence for a disfluency bias in non-native speech and to facilitate comparison to Bosker et al. (2014a), we repeated Experiment 1 (i.e., not Experiment 2) with non-native speech materials.

Method

Participants. Native Dutch participants (*N* = 38) with normal hearing were recruited from the Max Planck Institute's participant pool. Data from six participants were excluded due to technical failures (*n* = 3) or trouble with eye-tracking calibration (*n* = 3). In addition, at debriefing, two participants reported to have detected the particular disfluency distribution they were exposed to in the experiment (both from the Atypical Group); their data were also excluded. After exclusions, data from a total of 30 participants were left for analysis (25 females, 5 males; mean age = 22; range = 19–25), half of whom were assigned to the Typical Group, the other half to the Atypical Group.

Materials and design. For Experiment 3, we adapted the non-native speech materials from Bosker et al. (2014a). These involved the same types of recordings as the native materials but this time produced by a female non-native speaker of Dutch (L1 Romanian, LoR = 3.5 years, self-estimated CEFR level A1/A2, very limited experience using Dutch in daily life), with a strong foreign (Romanian) accent (average accentedness rating of 6.1 on a 9-point scale; cf. Bosker et al., 2014a). Again target fragments were excised and spliced onto one of three fluent sentence tokens and one of three disfluent sentence tokens. As a result, stimuli in Experiment 3 were identical to those in Experiment 1, except for the presence of a strong foreign accent.

Procedure. The procedure of Experiment 3 was identical to that of Experiment 1 (including the cover story, assigning participants to one of two groups, debriefing, etc.), except that this time participants were explicitly instructed that they would be presented with speech from a non-native speaker of Dutch. Although, once again, many participants noticed the relatively high disfluency incidence, only two participants explicitly reported (after the experiment) the (atypical) disfluency distribution they had been exposed to, and as a consequence were

excluded from analyses.

Results

Mouse clicks. Participants were perfectly accurate in their mouse clicks; no participant ever selected a competitor image over a target image. Before analyzing the reaction times, trials with outlier RTs (> 2SD above the mean; *n* = 101; < 3%) were excluded from analyses. Raw reaction times (RTs), calculated from target onset (in ms), from Experiment 3 are given in Fig. S3 in the Supplementary Material. Mean (SD) reaction times were: $HF_{Typical}$ = 1017 (272); $LF_{Typical}$ = 1098 (288); $HF_{Atypical}$ = 1045 (273); $LF_{Atypical}$ = 1126 (269); see also Fig. 8.

A Linear Mixed Effects Regression analysis tested the log-transformed RTs. This analysis involved the same structure as the one reported for Experiment 1, testing for effects of Group, Frequency, Trial Number, and all interactions. Statistical significance was assessed at the
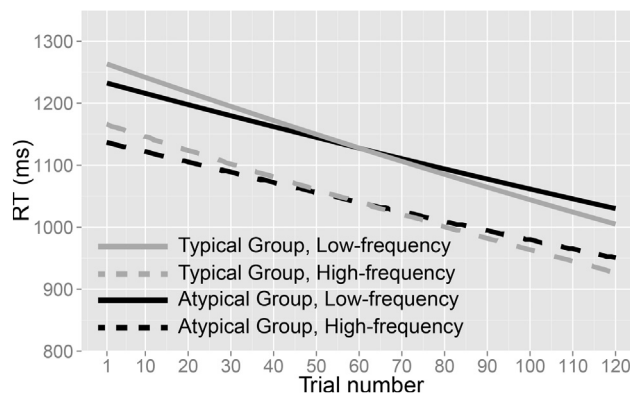

Fig. 8. Modelled reaction times (RTs) in ms from Experiment 3 (non-native speaker) as a function of the trial number, split by target frequency (solid lines show low-frequency targets, dashed lines show high-frequency targets) and group (Typical Group in gray, Atypical Group in black), as predicted by the Linear Mixed Effects Regression model.
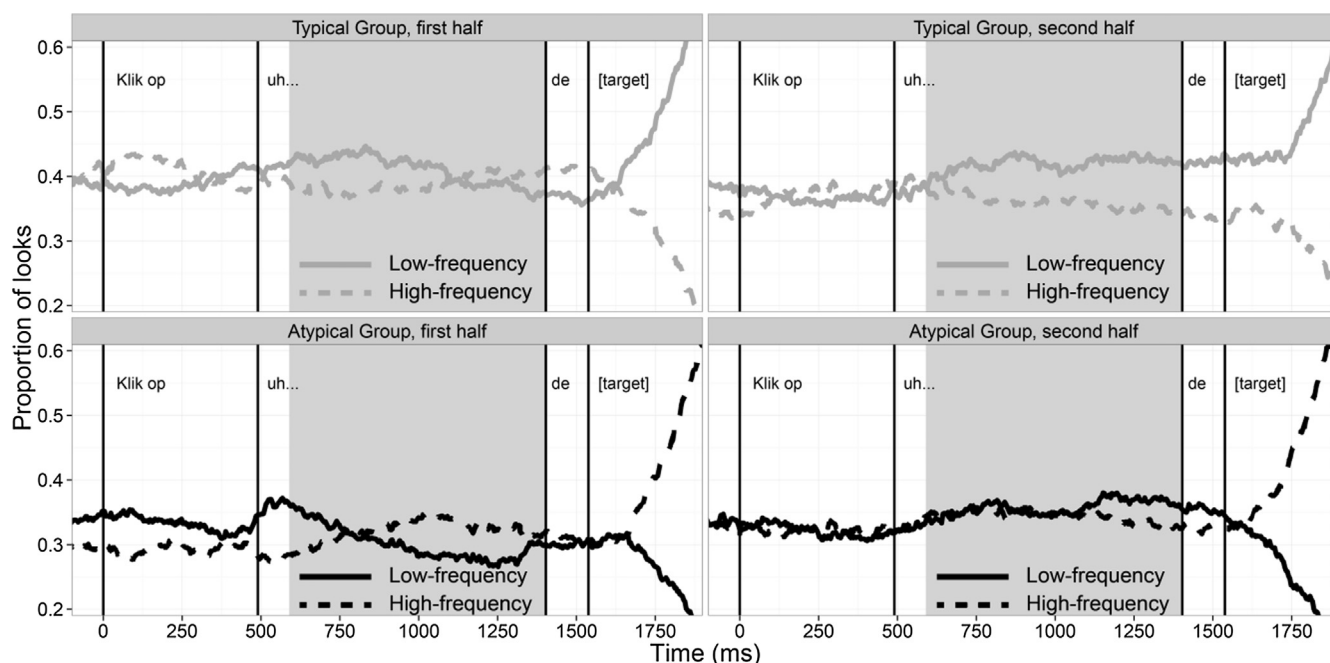
Fig. 9. Proportion of looks, in disfluent trials only, from Experiment 3 (non-native speaker) to high-frequency objects (dashed lines) and low-frequency objects (solid lines), separately for the two groups and the two halves of the experiment. Time in ms is calculated from sound onset. Vertical lines represent the (mean) onsets of words in the sentence. The gray rectangle indicates the time window of interest as identified for the statistical analyses. Note that the data are shown separately for the first and second half for illustration purposes only; all statistical analyses involved the predictor Trial Number.

0.05 significance level by checking whether effects had absolute *t*-values exceeding 2.

This model revealed an effect of Frequency ($\beta = 0.081$, $SE = 0.018$, $t = 4.410$; longer RTs for LF targets) and Trial Number ($\beta = -0.067$, $SE = 0.004$, $t = -16.260$; shorter RTs in later trials), similar as in Experiment 1. No overall difference between the two groups was found, and, in contrast to Experiment 1, no interaction between Group and Trial Number could be established ($\beta = 0.014$, $SE = 0.008$, $t = 1.800$). As such, there was no evidence for either Group showing a different order effect, even though, numerically, the Typical Group seemed to show a slightly steeper slope (cf. Fig. 8).

Eye fixations – Disfluent speech. Prior to the analyses, blinks and saccades were excluded from the data. Only fixations on the images themselves were coded as looks toward a particular image. The eye fixation data were downsampled to 250 Hz for simplicity.

Similar to Experiment 1, the fluent and disfluent condition were analyzed separately. In the disfluent condition, the same time window as previously was used, starting from *uh* onset + 100 ms up to article onset; see Fig. 9 (mean length: 813 ms).

We statistically tested the binomial fixations on low-frequency referents in the disfluent condition using a GLMM with identical structure as the one reported in Experiment 1. Note that this also entailed that the Typical Group was mapped onto the intercept, meaning that main effects should be interpreted with respect to the Typical Group only.

This GLMM revealed significant effects of Time ($\beta = -0.034$, $SE = 0.006$, $z = -6.040$, $p < 0.001$; initially, participants in the Typical Group started out looking away from the low-frequency referent over the course of the time window), Trial Number ($\beta = 0.022$, $SE = 0.006$, $z = 3.924$, $p < 0.001$; the proportion of looks to low-frequency referents gradually increased across the experiment for the Typical Group), and an interaction between Time and Trial Number ($\beta = 0.033$, $SE = 0.006$, $z = 5.896$, $p < 0.001$), indicating that participants in the Typical Group actually looked more and more at the low-frequency referent over the course of the time window of interest as the experiment progressed. These outcomes resemble those from Experiment 1, except that in Experiment 1 we did not observe a main effect of Time.

Interactions with the predictor Group revealed differential effects in the Atypical Group. A marginally significant interaction between Group and Time ($\beta = -0.013$, $SE = 0.008$, $z = -1.711$, $p = 0.087$) suggests that participants in the Atypical Group tended to look away from the low-frequency referent over the course of the time window. A significant interaction between Group and Trial Number ($\beta = 0.075$, $SE = 0.008$, $z = 9.572$, $p < 0.001$) showed that the proportion of looks to low-frequency referents *increased* (i.e., stabilized) as the experiment progressed in the Atypical Group, despite the fact that the disfluency *uh* only occurred before high-frequency referents in the Atypical Group. Finally, a three-way interaction between Group, Trial Number, and Time ($\beta = 0.039$, $SE = 0.008$, $z = 5.044$, $p < 0.001$) showed that this gradual *increase* in proportion of looks to low-frequency referents was even greater in later trials in the Atypical Group.

Finally, similar to Experiment 1, we statistically tested for early looking preferences in the time window from sound onset up to *uh* onset for each participant. We found no statistical evidence for early looking preferences for either HF or LF referents in either group.

Eye fixations – Fluent speech. In the fluent condition, the time window of interest started from sound onset + 100 ms up to article onset; see Fig. 10 (mean length: 313 ms).

We statistically tested participants' fixations on *high-frequency* referents in this time window by means of another GLMM with identical structure as used in Experiment 1. Note that the Typical Group was again mapped onto the intercept; simple effects should hence be interpreted with respect to the Typical Group only.

The model did not reveal any effect of Time ($p = 0.371$) but did show an effect of Trial Number ($\beta = 0.062$, $SE = 0.011$, $z = 5.385$, $p < 0.001$; the overall proportion of looks to high-frequency referents gradually increased across the experiment for the Typical Group). However, an interaction between Time and Trial Number ($\beta = -0.042$, $SE = 0.011$, $z = -3.694$, $p < 0.001$) demonstrated that the overall proportion of fixations to high-frequency referents *decreased* over the course of the experiment for the Typical Group. Finally, only one interaction with Group was found (Group * Trial Number: $\beta = 0.078$, $SE = 0.016$, $z = 4.823$, $p < 0.001$) showing that the proportion of looks to high-frequency referents *increased* as the experiment
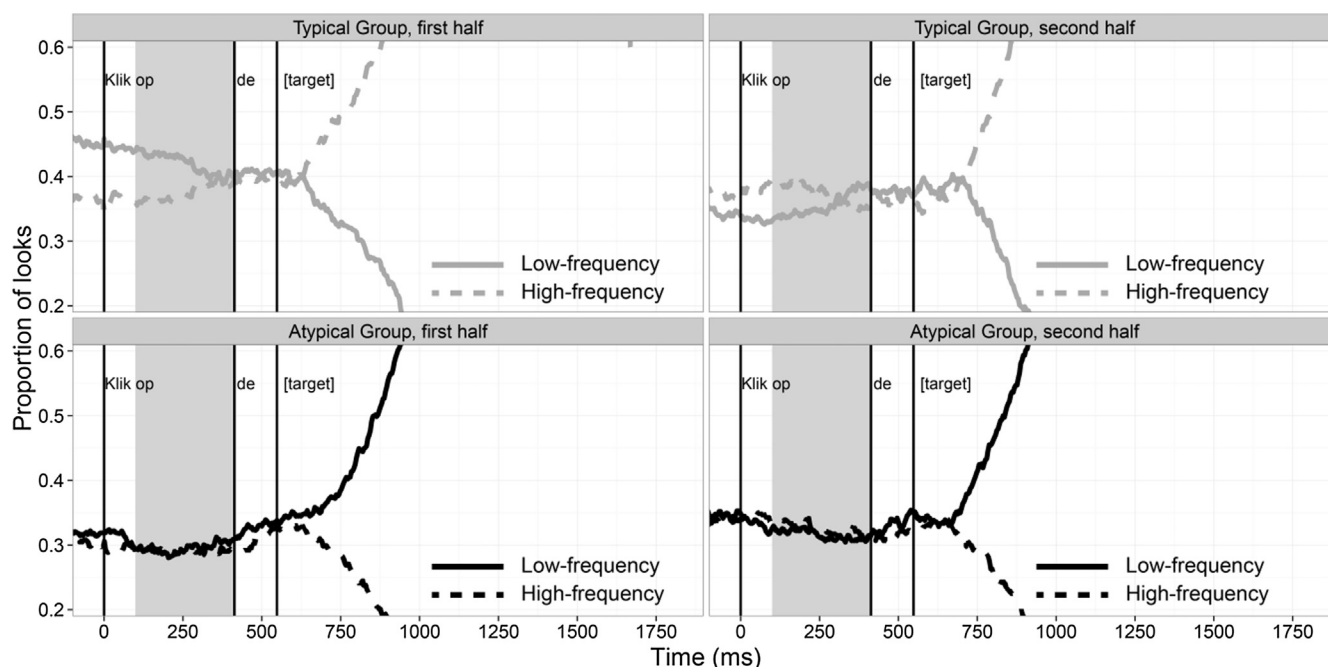
Fig. 10. Proportion of looks, in fluent trials only, from Experiment 3 (non-native speaker) to high-frequency objects (dashed lines) and low-frequency objects (solid lines), separately for the two groups and the two halves of the experiment. Time in ms is calculated from sound onset. Vertical lines represent the (mean) onsets of words in the sentence. The gray rectangle indicates the time window of interest as identified for the statistical analyses. Note that the data are shown separately for the first and second half for illustration purposes only; all statistical analyses involved the predictor Trial Number.

progressed in the Atypical Group.

Omnibus analysis

The primary difference in gaze patterns between Experiment 1 and Experiment 3 was that, in Experiment 1, the Typical Group anticipated low-frequency referents in disfluent trials, while the Atypical Group learnt to anticipate high-frequency referents in disfluent trials. By contrast, in Experiment 3, the Typical Group did learn to anticipate low-frequency referents in disfluent trials, but the Atypical Group did not adjust to the atypical disfluency distribution. In order to statistically assess this difference between experiments, an additional analysis was performed on the combined data from Experiment 1 and 3.

In this omnibus analysis, the crucial factor of interest was whether the three-way interaction between Group, Trial Number, and Time was significantly modulated by Experiment (i.e., a four-way interaction between Group, Trial Number, Time, and Experiment). A statistical comparison between the model including the four-way interaction and the model without the interaction effect (using the anova() function in R) revealed that including the four-way interaction significantly improved model quality ($\chi^2(1) = 127.31$, $p < 0.001$). The four-way interaction between Group, Trial Number, Time, and Experiment was reliable in the omnibus GLMM ($\beta = 0.120$, $SE = 0.011$, $z = 11.282$, $p < 0.001$). This suggests that comprehenders adapt to atypical disfluency distributions (i.e., learning to anticipate high-frequency referents upon hearing a disfluency) when listening to native speech, but not when listening to non-native speech.

Discussion

The data from Experiment 3 demonstrated differential patterns of distributional learning for the Typical vs. the Atypical Group. The Typical Group, although starting out with a slight preference for high-frequency referents in disfluent trials, showed a gradual increase in looks to low-frequency referents as the experiment progressed. This suggests that the Typical Group was able to adjust their predictive

strategies to the more typical disfluency distribution, even when produced by a non-native speaker. However, no evidence was found for anticipation of high-frequency referents in fluent trials for the Typical Group.

The Atypical Group did not show evidence for distributional learning. No evidence was found for anticipatory looks to *high-frequency* referents in disfluent trials, nor to *low-frequency* referents in fluent trials. Also no interaction with Group was observed in the RT analysis of Experiment 3. Thus, while for native speech we found the strongest evidence for distributional learning in the Atypical Group, no evidence was found for adaptation to a non-native atypical disfluency distribution.

General discussion

The present study demonstrated that listeners adapt to the experimental distributions of disfluencies in native and non-native speech. However, the distributional learning observed for native speech (Experiments 1 and 2) was qualitatively different from the distributional learning observed for non-native speech (Experiment 3).

In Experiment 1, we found that exposure to a 'typical' disfluency distribution (*uh*'s before low-frequency referents, no *uh*'s before high-frequency referents) led to a gradual increase in anticipatory fixations to low-frequency referents upon hearing the disfluency *uh*. Conversely, exposure to the opposite 'atypical' disfluency distribution (*uh*'s before high-frequency referents, no *uh*'s before low-frequency referents) led listeners to shift their initial disfluency-driven predictive strategies towards high-frequency referents. Additionally, Experiment 2 demonstrated that the same conclusions held when items were not repeated within an experimental session, removing the possibility for item-level priming. As such, Experiments 1 and 2 show empirical evidence for distributional learning in the online processing of disfluencies, providing support for the *associative account* of the disfluency bias. Associations between disfluency and the frequency of occurrence of following referents can be learnt and causally shape the disfluency bias, even within a single experimental session.

Of course, the 'typical' and 'atypical' distributions used in the present study represent rather extreme examples of disfluency distributions; they do not reflect the much more balanced disfluency distribution observed in spontaneously produced (native and non-native) speech (De Jong, 2016). Nevertheless, we think that the present findings obtained with these arguably unnatural distributions are highly relevant and carry implications for our understanding of the processing of more natural fluent and disfluent speech, based on three arguments.

First, even though the 'typical' and 'atypical' disfluency distributions were extreme, participants generally did not explicitly notice an association between (dis)fluency and target word characteristics (except for three participants in Experiment 2 and two participants in Experiment 3, all excluded from analyses). As such, the distributional learning observed here likely involves mechanisms of implicit learning that may well generalize to more natural conversation. Second, although the distributions used here may be argued to be unnatural, similar arguments can be raised against other previously used distributions in earlier studies on disfluency processing. These typically fully crossed the factor Fluency with the particular target manipulation at hand (e.g., disfluency was equally likely to occur before known vs. unknown objects; Arnold et al., 2007), which also deviates considerably from the natural disfluency distribution in native speech (i.e., biased towards more complex referents). Finally, although disfluencies occurring before high-frequency referents are unlikely (compared to occurring before low-frequency referents), they are not rare (De Jong, 2016). Based on these arguments, we propose that the present findings should be taken as a case in point: Listeners are demonstrated to be capable of implicitly tracking different disfluency distributions. Further research is required to investigate the stability and robustness of this type of distributional learning when it comes to more natural and more balanced disfluency distributions.

A related issue concerns the fact that the present evidence for distributional learning in disfluency processing was obtained using relatively simple two-image visual displays. The use of minimally complex visual arrays was motivated by (1) the desire to provide participants with as much opportunity as possible to adapt; and (2) wanting to stay as close as possible to the methods used in Thacker et al. (2018b) and Bosker et al. (2014a) to allow for comparison across studies. Whether similar distributional learning may be observed in more natural environments remains an open question. It is known that empirical evidence for prediction in language processing (typically measured as anticipatory eye movements in the visual world paradigm) is found as a function of the speech rate of the spoken stimuli, the amount of preview, the instructions, visual complexity, etc. (Ferreira, Foucart, & Engelhardt, 2013; Huettig & Guerra, 2019). Therefore, the present results show that listeners are capable of implicitly tracking different disfluency distributions in highly controlled lab environments, but future work may explore the limits of this behavior in arguably more natural settings.

One implication of finding distributional learning in disfluency processing is that adaptation to disfluency distributions likely also played a role in earlier studies on disfluency processing (e.g., Arnold et al., 2007; Barr & Seyfeddinipur, 2010; Heller et al., 2014). As explained above, these studies mostly used more balanced distributions, which also do not reflect the true natural disfluency distribution in spontaneous speech (De Jong, 2016). This may reduce the disfluency bias as listeners receive more and more exposure to the unnaturally balanced disfluency distribution (although few studies reported and/or tested for learning effects; see Introduction), leading to a potential underestimation of the true effect size of the disfluency bias in natural language comprehension. Future experiments should therefore include tests for learning effects.

Experiment 3 presented some constraints on a purely associative account by investigating non-native speech. When exposed to a 'typical' disfluency distribution produced by a non-native speaker, we also found evidence for distributional learning (similar to the Typical Group

in Experiment 1): listeners could learn to use the disfluencies produced by a non-native speaker to predict upcoming low-frequency referents. However, evidence for distributional learning in non-native disfluency processing was only found in the Typical Group. The Atypical Group in Experiment 3 did not show adaptation to the non-native atypical disfluency distribution, even though the disfluency distribution was identical to the one used in the Atypical Group in Experiment 1.

The differential tracking of disfluencies in native and non-native speech suggests, first, that non-native disfluencies can guide the disfluency bias *when presented in a native-like distribution*. The fact that the Typical Group in Experiment 3 could learn to anticipate low-frequency referents following non-native disfluencies shows that listeners do not abandon prediction altogether when listening to a non-native talker. As long as there is sufficient evidence of a native-like use of disfluencies in a non-native talker's utterances, listeners can adjust their predictive strategies to cleverly make use of the disfluency cues in non-native speech – extending the initial claims by Bosker et al. (2014a).

At the same time, listeners in the Atypical Group in Experiment 3 did not learn to predict high-frequency referents after hearing a non-native *uh*, despite the atypical disfluency distribution being as informative about the following referent as the typical distribution. This finding cannot be explained by a purely associative account of disfluency processing, since the associations between disfluency and referents' frequency of occurrence were equally strong in the two distributions and across experiments. Instead, the absence of evidence for learning from a non-native atypical disfluency distribution suggests that listeners also draw pragmatic inferences about the likely cause of disfluencies – in line with an *inferential account* of the disfluency bias. It could be that exposure to a few atypical disfluent instructions (i.e., hearing the non-native talker say *uh* before high-frequency words like "hand" or "car") led listeners to infer that the non-native talker at hand had considerable production difficulty retrieving even simple words in Dutch. Based on this inference, they may have taken the non-native disfluencies to not be predictive of the word to follow – in spite of the clear distributional cues indicating otherwise. As a result, they did not anticipate high-frequency referents after disfluencies and also did not adjust their predictive strategies as a function of the disfluency distribution they were exposed to.

Thus, the present study argues for a combination of *associative* and *inferential* processes to explain the disfluency bias in native and non-native speech. The experiments demonstrate that listeners track associations between disfluencies and following words and can adjust their anticipatory processing to these associations built up over the course of a single experimental session, in line with an associative account of the disfluency bias. At the same time, there are constraints on this distributional learning. Listeners can draw inferences about the cause of disfluencies which in turn may modulate the extent to which listeners adjust their predictive strategies, potentially accounting for the lack of distributional learning in the Atypical Group in Experiment 3.

A possible mechanism underlying distributional learning in disfluency processing may be prediction error. In predictive coding frameworks, learning depends on violations of expectations: the detection of differences between expected sensory input and actual sensation (Chang, Dell, & Bock, 2006; Ramscar, Dye, & McCauley, 2013). One specific prediction of a predictive coding framework is that disconfirmed predictions (in contrast to confirmations) lead to greater prediction error and hence to faster learning. This may explain why in our first two native speech experiments we observed the greatest adjustments (in RTs and looking behavior) and fastest learning in the Atypical Group, not in the Typical Group. Potentially, the contrast between participants' original expectations (*uh* typically precedes low-frequency referents) and the experienced atypical disfluency distribution (*uh* suddenly only precedes high-frequency referents) induced greater prediction error and thus most evidence of distributional learning.

The distributional learning observed in the present experiments

supports viewing the brain as a "prediction machine" (Clark, 2012), modifying perception and behavior so as to reduce error signals. Moreover, it shows that distributional learning in language comprehension is not limited to the linguistic content (i.e., phonology, syntax, semantics) of the communicative message; listeners even track such performance cues as disfluencies. Thus, the current results highlight the central role of prediction and adaptation in every aspect of spoken interaction (Pickering & Gambi, 2018). One implication following from this observation is that other metalinguistic, or even contextual, cues may also be actively tracked by listeners in online language comprehension, such as background noise (cf. McQueen & Huettig, 2012).

At the same time, the present outcomes demonstrate that the brain is no 'dumb' prediction machine. Distributional learning is sensitive to contextual cues that may modulate adaptation. Specifically, the same distributional cues were present in the atypical disfluency distributions in Experiments 1–3, yet only evidence for adaptation was found in Experiments 1 and 2. One parallel study to the present modulation of disfluency adaptation, driven by talker-identity, was carried out by Kraljic et al. (2008). They demonstrated modulation of phonetic adaptation induced by contextual factors, reporting that listeners do not adapt to a particular talker's lisp when there are incidental factors that explain the phonetic mispronunciation (e.g., a pen in the mouth; Kraljic et al., 2008). Phonetic adaptation has been successfully modeled using Bayesian inference and belief-updating frameworks (Kleinschmidt & Jaeger, 2015). In such frameworks, listeners are thought to not only use previous experience with cue distributions to predict upcoming material, but also to track situation-specific statistics, tuning their expectations based on top-down knowledge, with behavioral consequences. Thus, the parallels of the present findings on disfluency adaptation to earlier work on phonetic, syntactic, and semantic adaptation call for a unified, domain-general computational treatment of adaptation to linguistic, metalinguistic, and contextual cues in interactive communication.

Author contributions

Acknowledgements

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2019.02.006.

References

Altmann, G. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica, 137*, 190–200.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research, 32*, 25–36.

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say -thee uh- you're describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 914–930.

Arnold, J. E., & Tanenhaus, M. K. (2011). Disfluency effects in comprehension: How new information can become accessible. In E. Gibson, & N. Perlmutter (Eds.). *The processing and acquisition of reference* (pp. 197–217). Cambridge, MA: MIT Press.

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science, 15*, 578–582.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barr, D. J., & Seyfeddinipur, M. (2010). The role of fillers in listener attributions for speaker disfluency. *Language and Cognitive Processes, 25*, 441–455.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*, 123–147.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T. J. M., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*, 157–175. https://doi.org/10.1177/0265532212455394.

Bosker, H. R., Quené, H., Sanders, T. J. M., & De Jong, N. H. (2014a). Native 'um's elicit prediction of low-frequency referents, but non-native 'um's do not. *Journal of Memory and Language, 75*, 104–116. https://doi.org/10.1016/j.jml.2014.05.004.

Bosker, H. R., Quené, H., Sanders, T. J. M., & De Jong, N. H. (2014b). The perception of fluency in native and non-native speech. *Language Learning, 64*, 579–614. https://doi.org/10.1111/lang.12067.

Bosker, H. R., & Reinisch, E. (2015). Normalization for speechrate in native and nonnative speech. In M. Wolters, J. Livingstone, Beattie B, R. Smith, M. MacMahon, J. Stuart-Smith, & J. Scobbie (Eds.), Proceedings of the 18th international congress of phonetic sciences 2015 [ICPhS XVIII], Glasgow.

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language, 125*, 82–93.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234.

Clark, A. (2012). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*.

Corley, M., & Hartsuiker, R. J. (2011). Why um helps auditory word recognition: The temporal delay hypothesis. *PloS One, 6* e19792.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 113–132. https://doi.org/10.1515/iral-2016-9993.

Dell, G. S., & Chang, F. (2013). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*, 1–9.

Ferreira, F., Foucart, A., & Engelhardt, P. E. (2013). Language processing in the visual world: Effects of preview, visual complexity, and prediction. *Journal of Memory and Language, 69*(3), 165–182. https://doi.org/10.1016/j.jml.2013.06.001.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. e77661 *PLoS One, 8*(10), https://doi.org/10.1371/journal.pone.0077661.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34*, 709–738.

Grodner, D., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In N. J. Pearlmutter, & E. A. Gibson (Eds.). *The Processing and Acquisition of Reference* (pp. 239–272). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262015127.003.0010.

Hanulíková, Van Alphen, P. M., Van Goch, M. M., & Weber, A. (2012). When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing. *Journal of Cognitive Neuroscience, 24*, 878–887.

Hartsuiker, R. J., & Notebaert, L. (2010). Lexical access problems lead to disfluencies in speech. *Experimental Psychology, 57*, 169–177.

Heller, D., Arnold, J. E., Klein, N., & Tanenhaus, M. K. (2014). Inferring difficulty: Flexibility in the real-time processing of disfluency. *Language and Speech* Online Pre-Publication.

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research, 1706*, 196–208. https://doi.org/10.1016/j.brainres.2018.11.013.

Kidd, C., White, K. S., & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science, 14*, 925–934.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*, 148.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts how listeners adjust to speaker variability. *Psychological Science, 19*, 332–338.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. In Proceedings of the annual meeting of the cognitive science society (vol. 34).

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words

will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America, 131*, 509–517.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*, 475–494.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289–306.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204–238.

Orena, A. J., & White, K. S. (2015). I forget what that's called! children's online processing of disfluencies depends on speaker knowledge. *Child Development, 86*(6), 1701–1709. https://doi.org/10.1111/cdev.12421.

Owens, S. J., & Graham, S. A. (2016). Thee, uhh disfluency effect in preschoolers: A cue to discourse status. *British Journal of Developmental Psychology, 34*(3), 388–401. https://doi.org/10.1111/bjdp.12137.

Owens, S. J., Thacker, J. M., & Graham, S. A. (2018). Disfluencies signal reference to novel objects for adults but not children. *Journal of Child Language, 45*(3), 581–609. https://doi.org/10.1017/S0305000917000368.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*. https://doi.org/10.1037/bul0000158.

Quené, H., & Van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication, 52*, 911–918.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*, 413–425.

R Development Core Team. (2012). R: A language and environment for statistical computing [computer program].

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language

learning: The curious absence of mouses in adult speech. *Language, 89*(4), 760–793. https://doi.org/10.1353/lan.2013.0068.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294. 1926.

Severens, E., Lommel, S. V., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica, 119*, 159–187.

Thacker, J. M., Chambers, C. G., & Graham, S. A. (2018a). Five-year-olds' and adults' use of paralinguistic cues to overcome referential uncertainty. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.00143.

Thacker, J. M., Chambers, C. G., & Graham, S. A. (2018b). When it is apt to adapt: Flexible reasoning guides children's use of talker identity and disfluency cues. *Journal of Experimental Child Psychology, 167*, 314–327. https://doi.org/10.1016/j.jecp.2017. 11.008.

van Bergen, G., & Bosker, H. R. (2018). Linguistic expectation management in online discourse processing: An investigation of Dutch inderdaad "indeed" and eigenlijk "actually". *Journal of Memory and Language, 103*, 191–209. https://doi.org/10.1016/ j.jml.2018.08.004.

Van Berkum, J. J. A., Van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*, 580–591.

Van Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication, 35*, 103–113.

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication, 50*, 81–94.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language, 87*, 128–143. https://doi.org/10.1016/j.jml.2015.08.003.