Narrowed Expectancies under Degraded Speech: Revisiting the N400

Antje Strauß, Sonja A. Kotz, and Jonas Obleser

Abstract

■ Under adverse listening conditions, speech comprehension profits from the expectancies that listeners derive from the semantic context. However, the neurocognitive mechanisms of this semantic benefit are unclear: How are expectancies formed from context and adjusted as a sentence unfolds over time under various degrees of acoustic degradation? In an EEG study, we modified auditory signal degradation by applying noise-vocoding (severely degraded: four-band, moderately degraded: eight-band, and clear speech). Orthogonal to that, we manipulated the extent of expectancy: strong or weak semantic context (±*con*) and context-based typicality of the sentence-last word (high or low: ±*typ*). This allowed calculation of two distinct effects of expectancy on the N400 component of the evoked potential. The sentence-final N400 effect was taken as an index of the neural effort of auto-matic word-into-context integration; it varied in peak amplitude and latency with signal degradation and was not reliably observed in response to severely degraded speech. Under clear speech conditions in a strong context, typical and untypical sentence completions seemed to fulfill the neural prediction, as indicated by N400 reductions. In response to moderately degraded signal quality, however, the formed expectancies appeared more specific: Only typical (+*con* +*typ*), but not the less typical (+*con* −*typ*) context–word combinations led to a decrease in the N400 amplitude. The results show that adverse listening "narrows," rather than broadens, the expectancies about the perceived speech signal: limiting the perceptual evidence forces the neural system to rely on signal-driven expectancies, rather than more abstract expectancies, while a sentence unfolds over time. ■

INTRODUCTION

When hearing speech, listeners can use at least two streams of information: perceptual information provided by the speech signal itself, sometimes referred to as the "bottom–up" stream, and predictions or expectancies, denoted as the "top–down" stream. This top–down stream is, of course, commonly dependent on global discourse knowledge, but in this study it is used in a more specific sense of accumulated semantic context as the signal unfolds.

It is unclear how these two streams interact, particularly at the neural processing level. An intuitive assumption would be one of top–down expectancies becoming dominant whenever bottom–up perceptual evidence is ambiguous. Without doubt, top–down phenomena, where patchy or ambiguous perceptual evidence is filled in, are a powerful mechanism (in the visual domain: Tallon-Baudry & Bertrand, 1999; in the auditory domain: for example, Riecke, Esposito, Bonte, & Formisano, 2009; Sivonen, Maess, & Friederici, 2006). However, recent psycholinguistic and psychoacoustic research has emphasized that the opposite may also be true: Acoustic challenges have been shown to prompt listeners to focus on the bottom–up perceptual evidence, as opposed to mainly relying on top–down (contextual, i.e., semantic) cues (Mattys, Brooks, & Cooke, 2009).

Of the few studies on semantic cues in degraded speech that exists, most have operated with a unitary, simplified concept of "context." This, necessarily, confounded various linguistic aspects that might be differentially affected by speech degradation—only an experimental separation into various "levels" of context would allow investigation of whether the expectancy-forming neural mechanisms are differentially susceptible to degradation.

The current study attempts to study the specific interactions of degradation and expectancy formation at the neural level, using a simple and well-established marker, the N400 component of the ERP (Kutas & Hillyard, 1980; see below).

Semantic Context

As outlined above, expectancies can derive from various linguistic factors. Early experiments provided evidence that the recognition of a word occurs faster in a sentence context, compared with isolated or listed word presentations (Miller, Heise, & Lichten, 1951). Besides the syntactic structure that sentences provide (Miller & Isard, 1963), the semantic context of congruent or predictable

Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

sentences facilitates processing (Stanovich & West, 1983; Kutas & Hillyard, 1980; Kalikow, Stevens, & Elliott, 1977).

How can the benefits from semantic context be measured? In their seminal study, Kutas and Hillyard (1980) introduced the "cloze test" (originally developed by Taylor, 1953) to find a quantitative evaluation of sentence ending probability. In this test, participants have to complete sentences with the most likely word that comes to their mind, capturing implicit knowledge about contextual suitability. A number of studies have consistently replicated the benefit of high over low sentence ending probability (e.g., Obleser & Kotz, 2010; Friedrich & Kotz, 2007; van den Brink, Brown, & Hagoort, 2006; Van Petten & Luka, 2006; Connolly, Stewart, & Phillips, 1990). Jurafsky (2003) argued that one reason for this could be that people make overly crude distinctions between congruence and incongruence or high and low predictability. In fact, neither of these concepts is categorical but, rather, they operate on a continuum in natural languages. In the past, there was a lack of a priori criteria and measures of congruency and predictability to allow for the parametric variation of such concepts. As a result, many studies confined themselves to investigating effects of single word frequency on spoken word recognition (Cleland, Gaskell, Quinlan, & Tamminen, 2006; Benki, 2003; Luce & Pisoni, 1998; Howes, 1957) or—more complexly—to looking at the effects of bigram frequency (e.g., Ferrand et al., 2011).

During the last decade, the collection of huge text corpora and the establishment of computational tagging algorithms have made it possible to calculate several frequency-based interdependencies of words. This puts us in the position of being able to generate a continuum of context-based typicality, the probability of a word given some previous context, which not only respects the single lexical frequency, but also bigram probabilities and lexical class probabilities (Geyken, 2011; a psycholinguistic term for this being collocation). The sensitivity gained by quantifying the contextual relation within a sentence will be utilized in this study.

Neural Signatures of Context in Language Comprehension

A prominent component of the ERP in response to words, as measured by EEG, is the N400. This negativity, peaking at around 400 msec after word onset, is used as a neural indicator of context-based expectations and actual word input. Kutas and Hillyard (1980) reported the first observation of increased amplitude in response to an incongruent sentence ending word (for reviews, see Van Petten & Luka, 2012; Lau, Phillips, & Poeppel, 2008; Kutas & Federmeier, 2000). Van Petten and Kutas (1990) found a general positive shift of ERP amplitudes, that is, a reduction of the N400 the later a word appeared in an unfolding sentence. Halgren and colleagues (2002) showed that when an open-class content word appears in earlier sentence

positions, the brain activation in the N400 time range is less wide spread in left temporal cortices. Both studies interpreted their results as reflecting the insufficient amount of predictive context up to this point in the sentence. This could also explain the high sensitivity to (semantic) violation at sentence endings.

Besides effects of repetition, word frequency, and sentential context on the amplitude of the N400, Federmeier and Kutas (1999) also found an influence of categorical typicality. This allows differentiation of sentential semantic context from expected semantic features. An example of a sentence context in their study was, "They wanted to make the hotel look more like a tropical resort, so along the driveway they planted rows of…". "Palms" would be the highest cloze probability completion because, first, the context constrains the sentence ending to a tropical plant and, second, palms are prototypical representatives of tropical plants. Moreover, the authors found not only a reduced N400 in response to "palms," but a moderately reduced N400 in response to "pines," and the most pronounced N400 in response to "tulips," suggesting that palms and pines share more semantic features than palms and tulips. Both palms and pines belong to the same category "tree," but in the context of the tropics, palms are more typical than pines. However, the term categorical typicality does not describe the relationship between sentence constituents but rather relies on prototype theory and feature semantics, which is why it is hard to extend this to other word classes, such as verbs or adjectives.

Therefore, we looked for a measure of typicality based on collocation statistics that would capture the distribution of a word and its contextual co-occurrence probabilities. This would relate our findings back to sentential semantic constraints and not to the hierarchical organization of prototypes in the mental lexicon (although this hierarchy is, to some extent, context dependent, as D'Arcy, Connolly, Service, Hawco, & Houlihan, 2004, and Federmeier & Kutas, 1999, have shown). In short, the current study focuses not on the categorical typicality but on the sentence context-based typicality.

Semantic Benefits in Adverse Listening

Although a whole tradition of behavioral studies have laid the ground work for understanding cognitive processes in adverse listening conditions (Mattys et al., 2009; Davis & Johnsrude, 2003; Pichora-Fuller, 2003; Stickney & Assmann, 2001; Kalikow et al., 1977; Miller et al., 1951), only a few neuroimaging studies (e.g., McGettigan et al., 2012; Davis, Ford, Kherif, & Johnsrude, 2011; Obleser & Kotz, 2010; Obleser, Wise, Alex Dresner, & Scott, 2007) and EEG studies (e.g., Boulenger, Hoen, Jacquier, & Meunier, 2011; Obleser & Kotz, 2011; Romei, Wambacq, Besing, Koehnke, & Jerger, 2011; Aydelott, Dick, & Mills, 2006; Connolly, Phillips, Stewart, & Brake, 1992) have taken on the issue of semantic or expectancy benefits in degraded speech.

For example, Aydelott et al. (2006) contrasted natural with low-pass filtered speech signals and showed a reduced N400 effect in response to incongruent sentence-final words under acoustic degradation. Likewise, Obleser and Kotz (2011) used very simple German sentences, varying in cloze probability under three degradation levels, and found the cloze-driven N400 amplitude decreased linearly with more signal degradation. In an fMRI version of this paradigm, the same authors showed that the cortical extent of activation in the superior temporal cortex not only varied with degradation (better signals yielding stronger and more extended activations along the entire superior temporal gyrus and sulcus, STG/STS), but that this degradation effect was modulated by contextual predictability: For high-cloze sentences, the degradation effects were confined to areas within and surrounding primary auditory areas, in contrast to the wide-spread bilateral anterolateral STG/STS activation for low-cloze sentences. This hints to a narrowing or pruning of brain activity, dependent on good predictions and moderate signal quality (Obleser & Kotz, 2010).

The present study aimed to specify how expectancies from context are formed and adjusted over the time course of a sentence under various degrees of acoustic degrada-tion. We aimed to study this phenomenon by using an estab-lished, time-sensitive, and comparably simple-to-acquire neural parameter (the N400 component of the ERP).

The design crossed a threefold factor degradation with a threefold factor semantic expectancy, which combined "context" and "typicality" manipulations (Figure 1A). "Context" of a sentence-final keyword was manipulated, as in a large number of previous studies, via the preced-ing sentence context: highly constraining verbs often co-occur with fewer specific nouns than low constraining ones ($\pm con$). In the strong context, however, we addi-tionally varied what we refer to as the "typicality" of the sentence-final object: We distinguished between high and low frequency co-occurrences of the verb–object$_{[AKK]}$ rela-tion ($+con \pm typ$). These choices were validated using corpus analysis (collocations) and empirical cloze tests (see Methods).

We first hypothesized that the presence of any expec-tancy effect (i.e., context or typicality) in the N400 win-dow would depend on signal quality. We thus expected strongest N400 effects under clear speech conditions. Sec-ond, we expected the context manipulation to be more salient than the comparably subtle typicality manipulation.
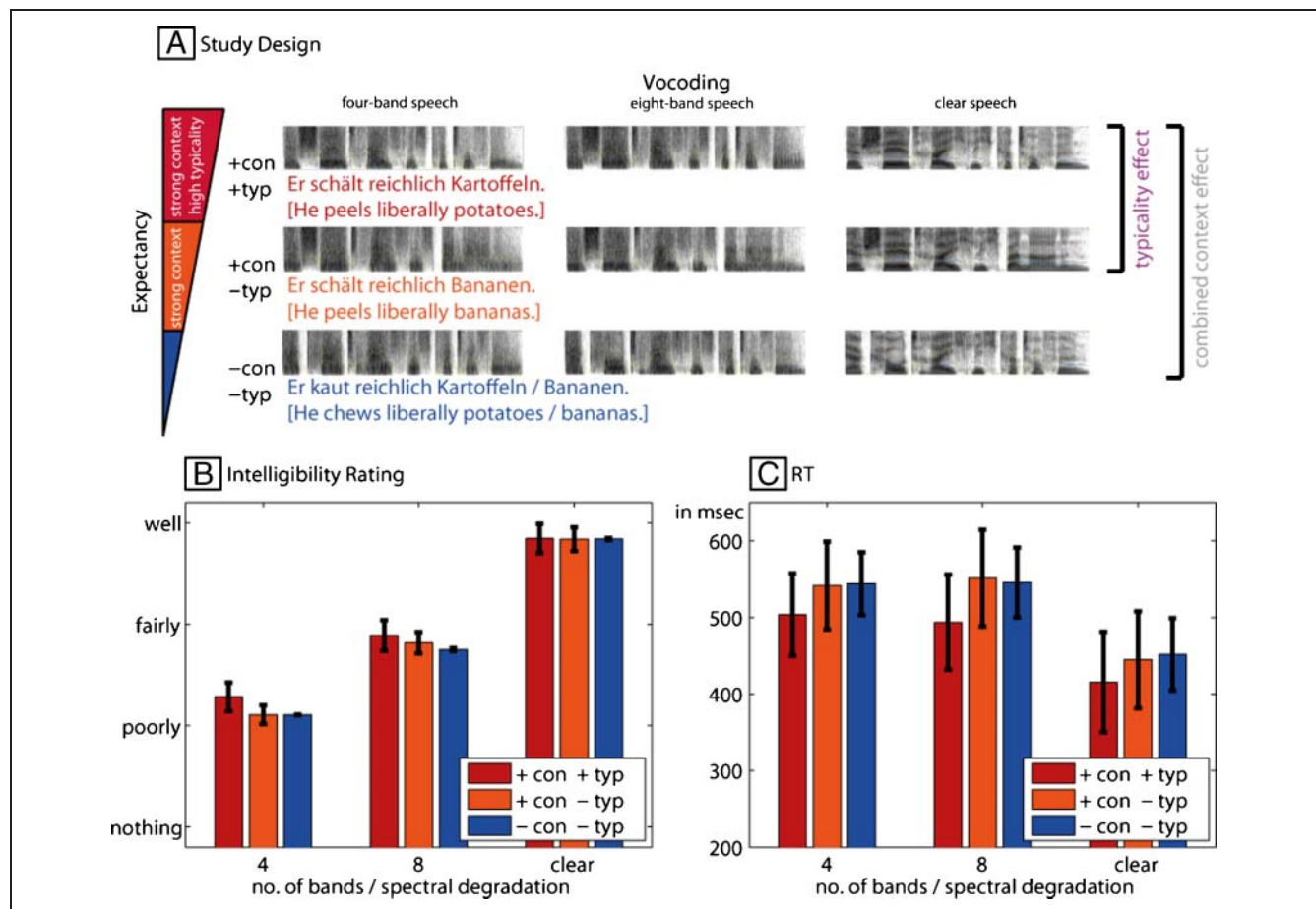

Figure 1. Study design and behavioral data. (A) Study design with factors Degradation and Expectancy and the two differences of interest. The effect of combined context is defined by condition $-con$ $-typ$ minus condition $+con$ $+typ$, and the effect of typicality only by $+con$ $-typ$ minus $+con$ $+typ$. (B and C) Behavioral results of the EEG experiment (mean ± 1 *SEM*).

Our third question, however, was the pivotal one: Would broad effects of "context" and more subtle effects of "typicality" behave the same under degraded and clear speech conditions? If acoustic degradation elicits a sharper or more specific adjustment of linguistic predictions as a sentence unfolds in time, then only the most typical word in a given context should match a formed prediction and effectively reduce the N400 effect.

METHODS

Participants

Twenty participants (13 women, 7 men; mean age = 25.7 years, SD = 2.64 years) took part in the auditory EEG experiment. All of them were native speakers of German and right-handed, with self-reported normal hearing abilities, no history of neurological or language-related problems, and no prior experience with vocoded speech. They gave their informed consent and received financial compensation for their participation. Thirty different participants were recruited for a behavioral pilot study. All procedures were approved by the ethics committee of the University of Leipzig.

Stimulus Material and Design

The study design was based upon three kinds of German sentences, varying in semantic context ($\pm con$) and context-based typicality ($\pm typ$), which will be outlined below in more detail. These were presented at three levels of speech signal degradation (severely degraded four-band speech, moderately degraded eight-band speech, and clear speech).

All sentences consisted of a pronoun ("er" masc. vs. "sie" fem.), a verb (in the present tense), an adverb, and an object. The neutral adverb, two or three syllables in length (e.g., "häufig" [often]), was inserted to temporally separate the two parts of interest (verb and object). Part of the material had already been used in previous studies on cloze probability (Obleser & Kotz, 2010, 2011; Gunter, Friederici, & Schriefers, 2000). For this study, the material was revised based on collocation statistics in the DWDS corpus (Digitales Wörterbuch der Deutschen Sprache: www.dwds.de, edited by Berlin-Brandenburgische Akademie der Wissenschaften). The corpus provides a measure of salience (Lin, 1998), on the basis of mutual information (MI; i.e., whether a combination of words co-occurs more often than chance). Different from the MI, however, the relative frequencies are not calculated over the whole text corpus but with respect to the syntactic relation (Geyken, 2011). This is especially relevant for the German language because the simple KWIC (Key Word In Context; also concordance), often used in English corpora, is inappropriate because of the less constrained word order and case syncretism in German (Geyken, 2011). To determine a meaningful measure of salience, word combinations have to co-occur at least four times in a specific syntagmatic relation in the DWDS corpus.

The semantic context was evoked by verbs with either strong or weak collocations: An ideal strongly determining context would have few co-occurring accusative objects and only one very frequent accusative object (e.g., "schält–Kartoffeln" [peels–potatoes]), whereas a weakly determining context would have a lot of equally (low) frequent alternative accusative objects (e.g., "kaut–Brot/Kaugummi/Fingernägel/Kartoffeln/etc." [chews–bread/chewing gum/fingernails/potatoes/etc.]). The context-based typicality was manipulated within the same semantic frame that each verb required. In the case of a strong context, this would be the contrast between the one very-high-frequency candidate and the one very-low-frequency candidate (but nevertheless co-occurring). In the case of a weak context, both candidates were selected to be equally probable.

Therefore, we defined "high-typical" in this study as a frequency tagging of the verb–object$_{[AKK]}$ relation greater than four (e.g., "schält–Kartoffeln," [peels–potatoes]) whereas "low-typical," that is, nonsalient combinations would be tagged fewer than four times in the DWDS corpus (e.g., "schält–Bananen," [peels–bananas]).

The less typical object of the semantic frame (i.e., "Bananen" [bananas] in the given example) always differed from the more typical target from the first phoneme on; where possible, the syllabic structure and stress pattern of the high-typical and low-typical objects were matched. In summary, 160 different sentences (40 themes × all 4 possible verb–object combinations) were created.

We also checked for single-word frequency: To estimate spoken word frequencies for all verbs and objects, we used the corpus of German movie subtitles, SUBTLEX, which has been shown to correlate better with lexical decision times than CELEX measures (Brysbaert et al., 2011). Word frequency = $\log_{10}$(item count + 2). High and low constraining verbs did not differ in their single word frequency (high: 1.49 ± 0.77, low: 1.69 ± 0.87), but typical objects were more frequent than untypical objects (typical: 2.41 ± 0.64, untypical: 1.94 ± 0.83). Note that, by manipulating the verb, we varied the occurrence probabilities of the objects so that single word frequencies would not play a role. More specific verbs are likely to be used in fewer contexts, which is why they are semantically constraining. Single word frequency of an object suitable for a highly constrained context can be high, which in our case condenses in a higher collocation frequency of the verb–object combination, that is, our main manipulation. This is because "Kartoffeln" [potatoes] and "schälen" [to peel] predict each other equally well, irrespective of what occurs first in a sentence.

All sentences were spoken by a phonetically trained female speaker and recordings were digitized at 44.1 kHz. Postediting included down sampling to 22.05 kHz, cutting at zero crossings, and RMS normalization. Additionally, each of the clear speech sentences was spectrally degraded using a Matlab-based noise-band vocoding algorithm

(70–9000 Hz, all vocoding-band envelopes smoothed with a 256-Hz zero-phase Butterworth low-pass filter). Levels of spectral degradation, that is, numbers of bands, were chosen according to a behavioral pilot study (see below).

Pilot Study

To select appropriate vocoding levels, we used a procedure for pretesting degraded speech stimuli (previously conducted by, e.g., Eisner, McGettigan, Faulkner, Rosen, & Scott, 2010; Obleser, Eisner, & Kotz, 2008; Obleser et al., 2007): Participants ($n = 30$; 15 women), who were not part of the EEG study described here, listened to all sentences at five different degradation levels (2-, 4-, 8-, 16-, and 32-band speech) and were instructed to type what they just heard. The first trial always picked a stimulus of the least degraded signal quality and was followed by a pseudorandomized order of sentences and degradation level. Feedback about correctness of response was provided only for the first 10 trials. Breaks were possible at participants' own discretion, resulting in an experimental duration of around 45 min. Accuracy was measured by taking the mean number of verb and object matches between each played sentence and the typed input.

For the EEG experiment, we identified eight-band speech as the critical condition, flanked by clear speech (no degradation) and four-band speech (hardly intelligible). These were selected because, first, it was at four and eight bands that the expectancy manipulation influenced comprehension. This was not, or only weakly, the case for 2-, 16-, and 32-band speech because of ceiling effects: four-band speech: +con +typ, e.g., "schälen–Kartoffeln" [peels–potatoes] = 63.8%; +con −typ, e.g., "schälen–Bananen" [peels–bananas] = 38.6%; −con +typ, e.g., "kauen–Kartoffeln" [chews–potatoes] = 41.3%; −con −typ, e.g., "kauen–Bananen" [chews–bananas] = 40.2%; eight-band speech: +con +typ, e.g., "schälen–Kartoffeln" [peels–potatoes] = 93%; +con −typ, e.g., "schälen–Bananen" [peels–bananas] = 88.6%; −con +typ, e.g., "kauen–Kartoffeln" [chews–potatoes] = 91.8%; −con −typ, e.g., "schälen–Bananen" [chews–bananas] = 73.6%. Second, eight-band speech yielded levels of comprehension that were approximately intermediate between four-band and highly intelligible 32-band speech.

EEG Acquisition

The EEG was recorded from 64 Ag-AgCl electrodes, positioned according to the extended 10–20 standard system on an elastic cap with a ground electrode mounted on the sternum. EOG was acquired bipolar at a horizontal (left and right eye corner) and a vertical (above and below left eye) line. All impedances were kept below 5 kΩ. Signals were referenced against the left mastoid and digitized on-line with a sampling rate of 500 Hz.

In an electrically shielded and sound-proof EEG cabin, participants were instructed to listen carefully to sen-

tences and rate them according to their intelligibility on a scale from 1 to 4, where 1 = *not at all comprehensible* and 4 = *perfectly understandable* (see Obleser & Weisz, 2012; Obleser et al., 2008; Davis & Johnsrude, 2003, for previous use of this rating task and close correspondence to actual comprehension). Responses were given via button press, and the button order was counterbalanced across participants. Seated comfortably in front of a computer screen, each participant listened to all 160 sentences at three degradation levels (in total 480 trials).

After each sentence, a question mark appeared on the screen prompting participants to give a rating. Subsequently, an eye symbol marked the time period for a blink break. Duration of the blink break and onset of the next sentence were jittered to avoid a contingent negative variation. Before the actual experimental trials, a short familiarization session was provided consisting of 10 trials (excluded from the analysis). Overall duration of the experimental procedure was about 1 hr.

Sentences were presented in a pseudorandomized order so that no more than two stimuli of the same signal quality were presented in succession and a clear speech sentence was heard in at least every fifth trial; also, the different expectancy manipulations belonging to one theme (i.e., one set of four sentences) were also presented at least 20 trials apart. The order of the clear speech and the degraded speech versions of one theme were changed across subjects to counteract facilitation through repetition. Nevertheless, contexts and objects were heard twice for each theme and, additionally, at three degradation levels, which is, in total, six times across the whole experiment. Despite careful randomization, it remains possible that repetition led to a reduction of the present effects. However, splitting trials would have prohibitively lowered the signal-to-noise ratio.

Individual electrode positions were determined after EEG recording with the Polhemus FASTRAK electromagnetic motion tracker (Polhemus, Colchester, VT, USA).

Data Analysis

Off-line preprocessing of data included rereferencing to linked mastoids, a finite impulse response high-pass filter at 0.03 Hz for drift removal, and automatic artifact rejection when EOG channels exceeded ±60 μV. Two different trigger points were used to average the EEG signal: For early ERP extraction, epochs of 3.2 sec (200 msec prestimulus baseline) were averaged, centered around the onset of sentences. For N400 analyses, the mean of the 2.2-sec epochs (200 msec prestimulus baseline), centered on the beginning of the sentences' final keywords, were considered.

Early ERP responses were analyzed at Cz. For the N100, two time windows of interest were defined: 50–100 msec and 100–150 msec, splitting the N100 into an early and a late time window to derive conclusions about the latency differences of the acoustic manipulation. For the P200, one time window from 150 to 300 msec was

identified. A repeated-measures ANOVA with the threefold factor of degradation (four-band, eight-band, and clear speech) was calculated for each time window separately.
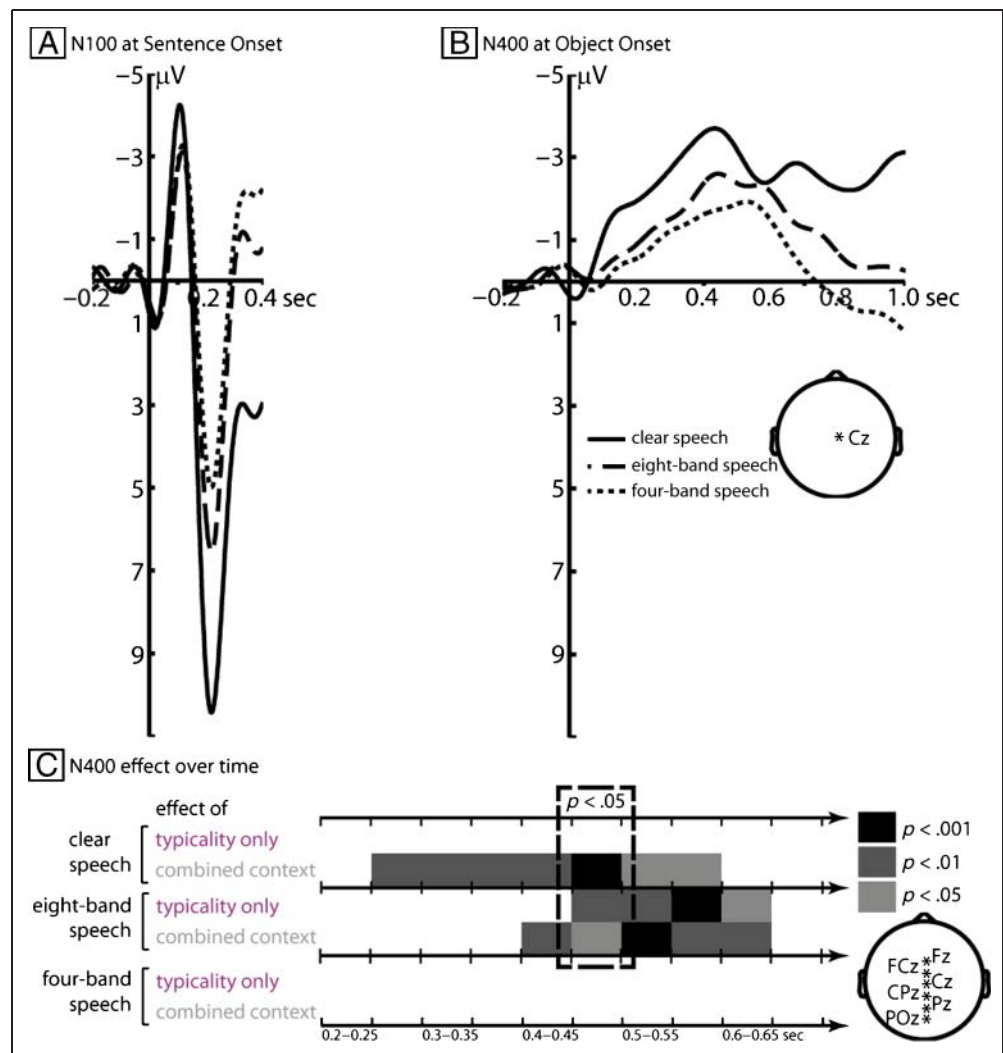
For the later ERPs associated with semantic processing, we merged both weak-context versions (e.g., "Er kaut reichlich Kartoffeln" [He chews liberally potatoes], "Er kaut reichlich Bananen" [He chews liberally bananas]) into one "weak-context, low-typicality" condition, because it is not possible to have a more or less typical completion in a low constrained context. To match the number of trials of this resultant $-con\ -typ$ control condition to the other two conditions, a random selection of trials was chosen. Thus, the final three conditions for data analysis each contained an equal number of trials. These final three conditions tested semantic context and typicality not in an orthogonal way, but rather, as a continuum of semantic expectation.

Generally, the N400 effect is defined as the difference between an expected, easy-to-integrate standard condition and some less expected, harder-to-integrate deviant condition. Therefore, we treated the $+con\ +typ$ condition

as standard, and calculated the difference waves for the "typicality-only effect" $[(+con\ -typ)-(+con\ +typ)]$ and the "combined context + typicality effect" $([(-con\ -typ)-(+con\ +typ)]$, for reasons of readability, this is referred to as the "combined context effect"; see Figure 1A for contrasts of interest). As the $-con\ -typ$ condition was a merged condition and consisted of $\pm typ$ nouns, the latter contrast combined a context effect with some portion of typicality effects. This is important to note and will be addressed when interpreting the results of the data.

In line with the N400 literature, we defined the scalp midline as the ROI and averaged the signal condition-wise across the electrodes Fz, FCz, Cz, CPz, Pz, POz. Confined to the midline ROI, we applied a time series analysis on the difference waves of the combined context effect and the typicality-only effect in all three degradation levels separately. By taking the mean over 50 msec time windows, we calculated 10 successive t tests against zero from 200 to 700 msec after object onset (Figure 2C) and corrected for multiple comparisons using the false discovery rate (FDR). To describe the differential N400 effects in

Figure 2. Grand-averaged ERP responses. (A) N1–P2 complex at sentence onset, by degradation level, shown at electrode Cz. (B) N400 at object onset, by degradation level, shown at electrode Cz. (C) Statistical analysis in 50 msec time windows after object onset. t Test of N400 difference waves against zero. Note that in four-band speech no robust N400 effect is detectable independent of condition. Also note that, in clear speech, no typicality only effect occurs whereas the context manipulation is robust and long-lasting. p Values shown survive FDR correction.

response to clear versus medium degraded speech, 2 × 2 repeated-measures ANOVAs with factors Degradation (eight-band, clear speech) and Expectancy Effect (typicality-only effect, combined context + typicality effect) were applied, first, on the N400 peak latencies which were extracted between 300 and 600 msec after object onset and, second, on the amplitude over a time window of 450–500 msec after object onset.

For the behavioral measures (RTs and accuracy), we performed a 3 × 3 repeated-measures ANOVA with factors Degradation (four-band, eight-band, and clear speech) and Expectancy (+con +typ, +con −typ, and −con −typ). p values were always acquired with Greenhouse–Geisser-corrected degrees of freedom. Nevertheless, degrees of freedoms are reported uncorrected throughout the manuscript for readability purposes. Where indicated by a significant interaction, further post hoc ANOVAs and t tests were calculated. Post hoc tests were corrected for multiple comparisons using FDR.

RESULTS

Intelligibility Rating and RT

Results of the intelligibility rating analysis show the two main effects of Degradation, $F(2, 38) = 51.58, p < .001$, and Expectancy, $F(2, 38) = 30.06, p < .001$, and a significant interaction of both factors, $F(4, 76) = 12.31, p < .001$ (Figure 1B). For each vocoding level, sentence types differed significantly (clear speech: $F(2, 38) = 3.44$, ns after FDR correction; eight-band speech: $F(2, 38) = 17.09, p < .001$; four-band speech: $F(2, 38) = 21.19, p < .001$).

At eight-band speech, intelligibility ratings linearly increased with semantic expectancy, that is, there was not only a benefit of context (+con typ vs. −con −typ: $t(19) = 3.55, p < .01$; +con +typ vs. −con −typ: $t(19) = 4.85, p < .001$), but also of typicality (+con +typ vs. +con −typ: $t(19) = 3.15, p < .01$). At four-band speech, only the strong-context, high-typicality condition differed from the other sentences and received markedly higher intelligibility ratings (+con +typ vs. +con −typ, $t(19) = 4.59, p < .001$, and +con +typ vs. −con −typ, $t(19) = 5.54, p < .001$).

As Figure 1C suggests, RTs showed main effects of Degradation, $F(2, 38) = 18.45, p < .001$, and of Expectancy, $F(2, 38) = 18.19, p < .001$, but no interaction ($F < 1$). These main effects are founded in faster responses under clear speech conditions (clear vs. eight-band speech: $t(19) = −5.04, p < .001$, clear vs. four-band speech: $t(19) = −5.25, p < .001$), and faster button presses for strong-context, high-typicality sentences compared with the other sentences (+con +typ vs. +con −typ: $t(19) = −5.35, p < .001$, +con +typ vs. −con −typ: $t(19) = −5.02, p < .001$, +con −typ vs. −con −typ: $t(19) = −0.14, ns$).

ERPs to Sentence Onset (N100/P200)

To assess the effects of degradation, disregarding the sentence-level expectancy manipulation, we first analyzed the evoked potential in response to sound (i.e., sentence) onset. Results are shown in Figure 2A. The N100 has a steeper slope and greater negative amplitude for clear speech than for degraded speech (at Cz 50–100 msec: main effect of Degradation $F(2, 38) = 4.97, p < .05$; clear vs. eight-band: $t(19) = −2.72, p < .05$; clear vs. four-band: $t(19) = −3.23, p < .01$; eight-band vs. four-band: $t(19) = 0.17, ns$; at Cz 100–150 msec: main effect of Degradation $F(2, 38) = 3.67, p < .05$; clear vs. eight-band: $t(19) = −2.38, p < .05$; clear vs. four-band: $t(19) = −1.67, ns$; eight-band vs. four-band: $t(19) = 1.13, ns$). The P200 shows a stepwise amplitude reduction depending on degradation level, with the highest amplitude in response to clear speech and the lowest in response to four-band speech (at Cz 150–300 msec: main effect of Degradation $F(2, 38) = 56.67, p < .001$; clear vs. eight-band: $t(19) = 7.39, p < .001$; clear vs. four-band: $t(19) = 9.91, p < .001$; eight-band vs. four-band: $t(19) = 2.64, p < .05$).

ERPs to Sentence-final Word (N400)

Our main hypotheses were focused on the N400 component of the evoked potential elicited at the sentence-final object. The typicality and combined context effects in the N400 were calculated as difference potentials to the +con +typ (standard) condition (see Methods). We began the N400 analysis by a series of 10 t tests over the midline ROI with a window length of 50 msec from 200 to 700 msec (testing the N400 difference waves against zero; Figure 2C). Only p values that survived FDR correction are shown.

Confirming our hypothesis, this analysis revealed that a weak N400-like effect in response to four-band speech (Figure 2C) was not significantly and consistently different from zero (see Figure 2C for details). For the typicality-only effect in four-band speech, there was no p value below .05, not even in one time window (the closest was found from 600 to 650 msec, $t(19) = −2.06, p = .053$), and for the combined context effect, three time windows differed from zero (the highest t value was from 550 to 600 msec, $t(19) = −2.92, p < .01$), but they did not survive the FDR correction. Hence, all further analyses reported here focuses on eight-band and clear speech only. Second, there was no consistent N400 typicality-only effect in response to clear speech, but the N400 typicality-only effect was strong and long-lasting in response to moderate degradation (eight-band speech; Figure 2C).

Generally, the N400 time window appeared to be delayed in response to eight-band speech. This was corroborated by a significant difference in N400 peak latency: For each subject and condition, the peak latencies of the N400 difference waves between 300 and 600 msec post word onset were extracted. A 2 × 2 repeated-measures ANOVA with factors degradation level (eight-band and clear speech) and expectancy difference (typicality and combined context + typicality) showed the N400 to peak
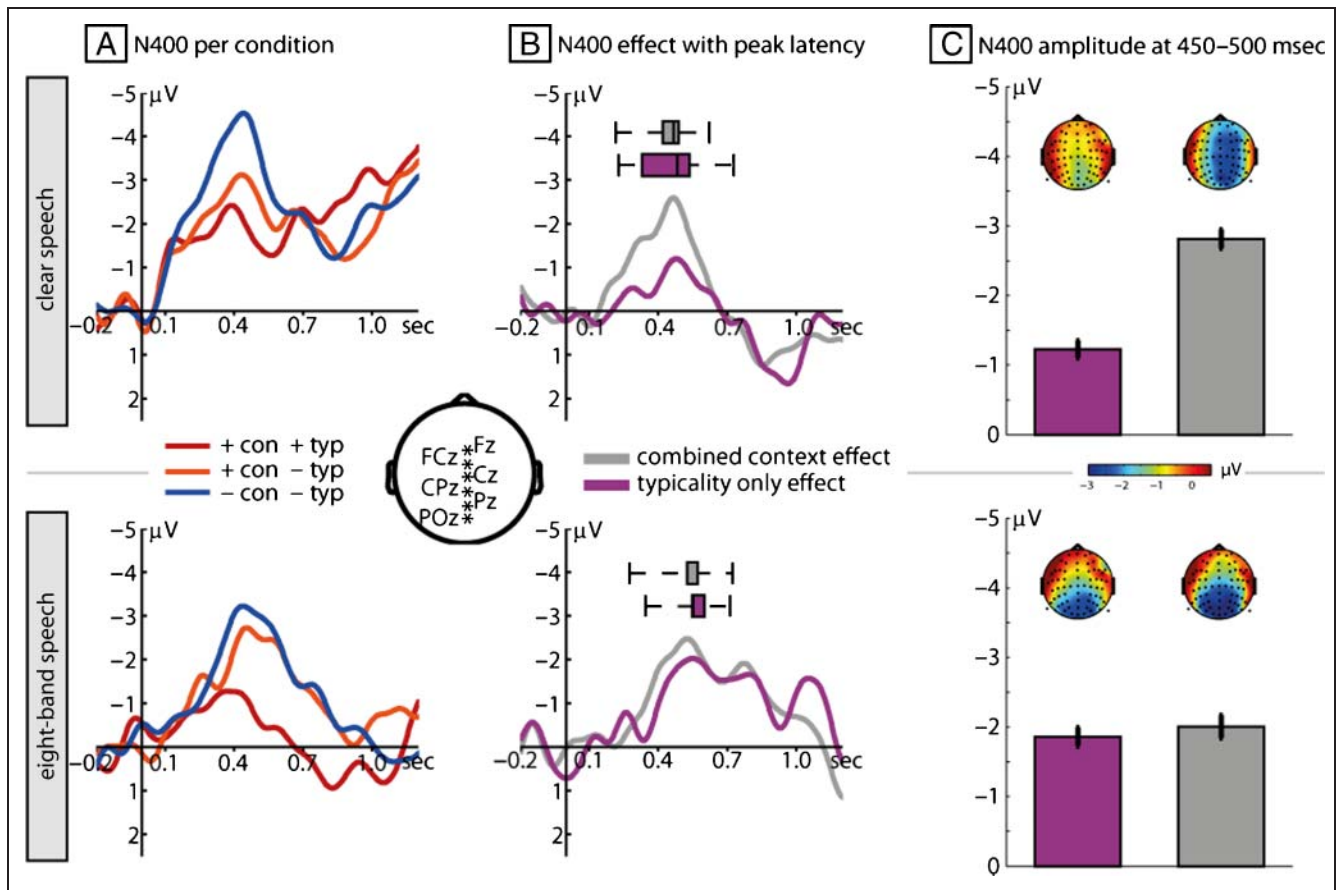
Figure 3. Differential N400 effects for clear and degraded speech. (A) Grand-averaged ERP responses at midline electrodes by expectancy level after object onset. (B) Difference waves of the ERPs and their N400 peak latency distributions (as depicted by the box plots). Note the significant delay of N400 peaks in eight-band speech. (C) Bar graphs of the N400 effects for combined context and typicality only at 450–500 msec after object onset with corresponding topographies.

around 78 msec earlier in response to clear speech (average peaks around 458 and 460 msec, respectively) than in response to eight-band speech (average peaks around 553 and 522 msec, respectively; $F(1, 19) = 8.28$, $p < .01$; Figure 3B).

A time window from 450 to 500 msec, which covers the N400 amplitudes regardless of the delay (as indicated by the dashed box in Figure 2C), was chosen and a 2 × 2 repeated-measures ANOVA with factors of degradation level (eight-band and clear speech) and expectancy difference (typicality and combined context + typicality) was calculated (Figure 3). We found a significant interaction, $F(1, 19) = 5.91$, $p < .05$. Post hoc t tests confirmed that the N400 effects of typicality and context differed significantly in response to clear speech (i.e., only the low context condition evoked an N400, while the typicality manipulation did not; $t(19) = 2.79$, $p < .05$), whereas in response to eight-band speech, there was no significant difference in the strength of the effects ($t(19) < 1$, ns; Figure 3C).

To summarize, speech degradation reduced the amplitude of the early, as well as the later, ERP responses, which is consistent with the stepwise decrease in the behavioral intelligibility ratings. Furthermore, signal deg-

radation not only delayed the N400 component but also interacted with the expectancy manipulation.

DISCUSSION

The goal of this study was to specify how expectancies may be formed from context and adjusted as a sentence unfolds over time, under various degrees of acoustic degradation. The central research question concerned how broad effects of context and more subtle effects of typicality may influence a neural marker of effortful integration, the N400, under degraded speech conditions. In contrast to a general broadening of semantic predictions under degradation, we hypothesized that acoustic degradation would, instead, elicit a sharpening and more narrow adjusting of linguistic predictions: Only the most typical word in a given context should match a formed expectancy and effectively reduce the N400 effect.

First, the occurrence of an N400 effect depended on the extent of signal degradation. There was no semantic modulation of the N400 in the severely degraded (four-band) speech condition, suggesting that fast linguistic processes were effectively hindered. In the moderately degraded

(eight-band) condition, the N400 amplitude was attenuated and the peak was significantly delayed for ~78 msec; this is in line with previous studies (Obleser & Kotz, 2011; D'Arcy, Service, Connolly, & Hawco, 2005; Holcomb, 1993; Connolly et al., 1992).

Second, the N400 reflected fine semantic differentiations of the context strength of a sentence and the typicality of a particular word in this context (Federmeier & Kutas, 1999; Connolly et al., 1992; Connolly & Phillips, 1994; Kutas & Hillyard, 1984). The combined context effect was generally more pronounced than the typicality-only effect and showed the known posterior-central scalp topography.

In addition, however, we found an expectancy differentiation in the N400 that was dependent on signal degradation: In the clear speech condition, a strong-context, low-typicality object appeared to be compatible with the predictions formed by the context, and the neural effort of integration (as reflected by the N400) was low in amplitude (Figure 3C; see also schematic display in Figure 4A). Note that unlike previous studies with similar manipulations (Desroches, Newman, & Joanisse, 2009; Newman & Connolly, 2009; Connolly & Phillips, 1994), we were unable to show N200- or PMN-like effects. This might be due to our task, which guided participants to focus on semantic rather than segmental information. In the moderate degradation condition (eight-band speech), however, the same strong-context, low-typical word triggered a pronounced N400 response that was statistically indistinguishable from the response to an unpredictable (i.e., weak-context, low-typical) word.

The topography of the N400 combined context effect at 450–500 msec was more frontally distributed in re-sponse to clear speech than degraded speech. A tentative explanation could be that, for clear speech, N400 sources might be more anterior and widespread than for acoustically degraded speech. In relation to spatially more precise functional MRI work on expectancies under degraded speech conditions, Obleser and Kotz (2010) found that the anterior STS/STG showed a linear increase of activation with a more intelligible signal. Interestingly, the same study generally reported more spatially constrained intelligibility activations in response to highly predictable sentences. A more widespread N400 in response to degraded speech was also reported recently (Romei et al., 2011), suggesting that additional attention or working memory processes are needed in adverse listening conditions. In contrast to our manipulation, however, this study used isolated words without sentential context and investigated the N400 not in response to the last word, but the intermediate word in a list of three.

Note that our manipulation (i.e., restricting the relevant semantic context to a single word: the preceding verb) also allows interpretations in terms of lexical priming. The adjectives inserted between the verb and target object were included to minimize this possibility. Nevertheless, associative priming may be observed even when an intervening item is presented (Joordens & Besner, 1992). Thus, the current results may not differentiate between sentential and lexical semantics but, instead, hint to the interesting fact that, even though participants focused on semantic information because of the task, they were utilizing it differently under degraded speech conditions. Somewhat counterintuitively, Mattys et al. (2009) showed that listeners in adverse hearing situations tend to rely less upon
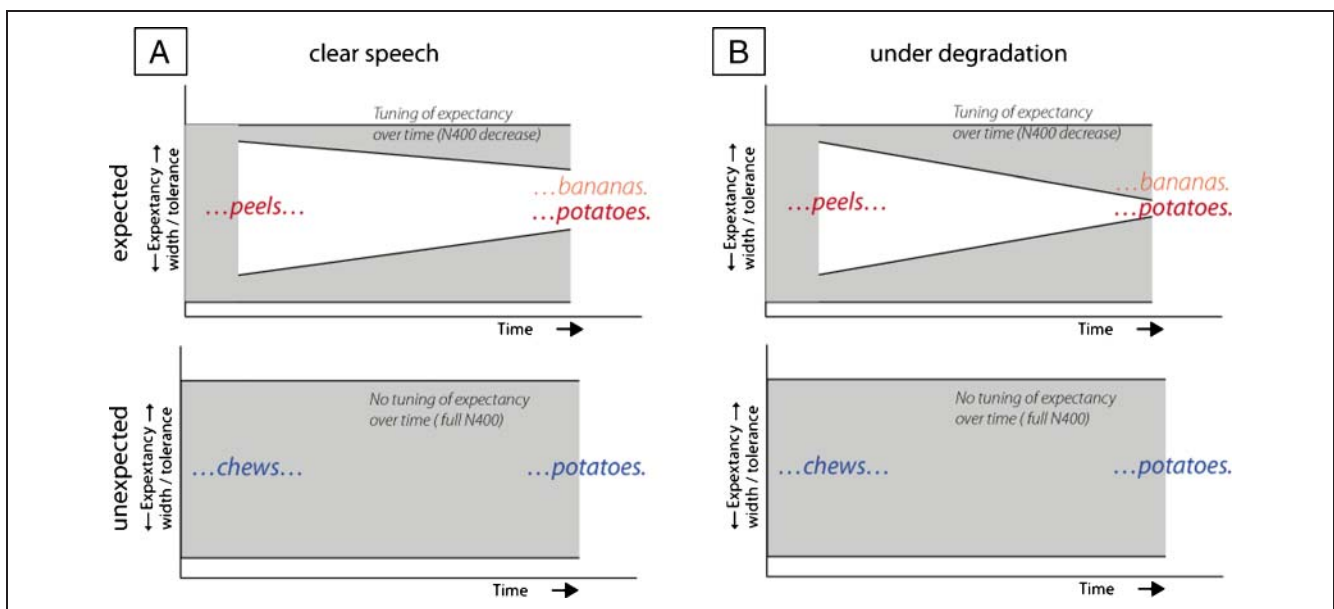

Figure 4. Model of expectancy searchlight. (A) Fast processing and context abstraction in clear speech, which can be characterized as a more tolerant "searchlight" process. (B) Fast recognition but decelerated context abstraction in moderate degradation (i.e., a narrowing of the expectancy searchlight). (A and B, bottom panels) No expectancy searchlight is formed when low context is provided independent of signal quality. See Discussion for details.

lexical-semantic cues and more on acoustic-phonetic detail. These results suggest that perceptual load might have narrowed the expectancy to an acoustic-phonetic focus, that is, unexpected segmental information could not be compensated by top–down knowledge (see section on "Prediction capacities and other cognitive resources" below).

The fact that less typical objects also had lower word frequency (see Methods) allows an alternative interpretation: comprehension and integration of lower-frequency words, in general, might benefit from strong contextual constraints in clear speech, but not in degraded speech. Note, however, that this interpretation would be also consistent with a basic conjecture of this study—namely, that the effects of contextual constraint on target processing differ depending upon the intelligibility of the acoustic signal and the probability of the target.

Overall, the present findings confirm that a degraded context is, in absolute terms, less effective at activating compatible semantic features than clear speech (cf. Aydelott, Baer-Henney, Trzaskowski, Leech, & Dick, 2012), such that contextual facilitation of unexpected but semantically consistent words is reduced. However, the results indicate that constraint-based expectancies that favor high-probability completions are relatively maintained under moderate perceptual degradation; consistent with previous findings that listeners' use of semantic cues in strongly biasing sentence contexts is relatively robust in adverse conditions (e.g., Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984; Kalikow et al., 1977).

N400 and Behavioral Responses: Fast versus Delayed Processes

A remaining open question is how to reconcile behavioral effects in the intelligibility ratings and brain effects in the N400 time range.

The absence of a semantic modulation of the N400, as the current data show in four-band speech, might be due to the lack of fast mapping capacities under poor acoustics. However, an intelligibility gain of strong-context, high-typical sentences in the ensuing behavioral response was still present. This suggests that the time it took for participants to generate a behavioral response allowed for retrospective semantic analysis of the degraded signal and affected intelligibility ratings for these sentences.

At intermediate signal degradation (eight-band speech), one could argue that the N400 is equally sensitive to different expectancy manipulations, whereas the behavioral data show a stepwise increase of intelligibility with growing expectation. Also, the N400 response is generally delayed under degraded speech compared with clear speech conditions. The benefit in performance for strong context, low typical sentences against unrelated sentences again suggests some later recovery, if at least some expectations could be formed.

This indicates that, under intermediate degradation, fast recognition and integration processes are possible (sensitivity to semantic expectations, i.e., reduced N400 in response to strong-context, high-typical words), but that they are still delayed when recognition and integration have to be based on less typical words (enhanced N400 in response to strong-context, low-typicality words). Put differently, an "expectancy searchlight" can be formed, based on sufficient perceptual evidence, but it will be narrowed because of limited cognitive resources (Figure 4B).

Finally, under clear speech conditions, we found an N400 combined context effect that was absent in the behavioral data. The N400, therefore, seems to reflect successful, albeit effortful, comprehension. Figure 4A displays it as a liberal expectancy searchlight where less thorough sentence processing in clear speech results from fast cue integration and context abstraction.

As Figure 4 summarizes, we suggest a tentative interpretation of our main findings in terms of an "expectancy searchlight." If listening conditions are ideal, expectancies are more liberal and the "searchlight" in a strong context is focused, but tolerant. The clear speech N400 effects were reduced in the strong context, irrespective of low or high typicality, compared with weak-context sentences. When dealing with acoustic limitations, however, this searchlight is narrowed down, and only the most typical sentence ending is facilitated in this case (Figure 4B).

Prediction Capacities and Other Cognitive Resources

The present data deliver important evidence for a trade-off between acoustics and semantics: First, the results of the N100–P200 complex at sentence onset suggest familiarization and categorization difficulties with degraded signals. We found the highest N100 amplitude in response to clear speech, and, in line with Obleser and Kotz (2011), no significant difference between four-band and eight-band speech. Moreover, Obleser and Kotz (2011) found the N100 response to be most pronounced in the 1-band speech condition, a highly unintelligible signal. Further testing indicated that the N100 amplitude has a u-shaped relation to speech intelligibility (Obleser et al., in preparation). The N100 is thought to index an initial allocation of resources and formation of a sensory memory trace (e.g., Schroeger, Trevaniemi, & Huotilainen, 2003), but it is unclear whether higher familiarity and easier categorization, as in clear speech, should lead to an increased or reduced N100. The current data, together with previous observations, suggest that the measured N100 amplitude is under the joint influence of low-level acoustic factors, such as perceived loudness and spectral resolution, and cognitive factors, such as familiarity. Furthermore, we observed the strongest P200 amplitude in response to clear speech and the weakest in response to four-band speech. Paulmann, Ott, and Kotz (2011) related their differential P200 responses to salient acoustic features (e.g., pitch,

voice quality, and loudness) of a stimulus. Less spectral information and, thus, reduced saliency of important acoustic features may lead to greater variance in the neural processing due to wide spread resource allocation for processing, which condenses in a reduced time-locked ERP response.

Second, all N400-like processes in response to eight-band speech were delayed in time, as shown by a significant difference in N400 peak latency that was driven by degradation. Also, RTs were longer in response to degraded sentences (Figure 1C). This is compatible with results by D'Arcy et al. (2005), who reported a reduced and delayed (~51 msec) N400 response to incongruent sentence ending words when working memory load was increased. More directly, evidence on the detrimental effects of speech degradation on working memory processes has accumulated (e.g., Obleser, Wöstmann, Hellbernd, Wilsch, & Maess, 2012; Piquado, Cousins, Wingfield, & Miller, 2010; Rabbitt, 1968).

To accomplish rapid speech comprehension in everyday communication, a language-familiarized listener constantly predicts forthcoming linguistic input (Gagnepain, Henson, & Davis, 2012). The adjustment of these predictions may be partly explained by psycholinguistic models that describe auditory language comprehension as a trade-off between perceptual evidence and other cognitive resources.

Norris and McQueen's (2008) Shortlist B model, for example, takes into account perceptual ambiguities and their interaction with word frequency. Recall that the conditional probability of the target word in a given context varied ("typicality"; [peel … potatoes] vs. [peel … bananas]). Consequently, an explanation arising from the Shortlist B model would be the following: Under clear speech, probability or typicality differences would be assumed to play only a negligible role, because all words would be correctly identified, and performance would approach ceiling. However, when the perceptual information is sparse (i.e., under degraded speech), the listener will have to resort to established word probabilities, and these probabilities would also affect the neural processes reflected by the N400. Thus, from such a cognitive psychology angle, it would be expected that acoustic degradation would narrow the range of lexical items that an automatic neural integration mechanism will pre-activate (and that will, hence, elicit only a small N400 amplitude; Figure 4).

Another interpretation of the observed adjusting of the range of expected words would be that perceptual load (i.e., the resources used for effortful processing of the signal itself) limits the resources a listener has available for forming predictions as the sentence unfolds. In this case, word probabilities would always be used (as they would under clear speech conditions) but are less accessible in adverse listening conditions because of shared resources (auditory and lexical analysis). Therefore, only the most probable ones would be preactivated.

Thus, in a Shortlist B framework, probabilities are used as an active compensation, whereas in a capacity limitation framework, shared cognitive capacities inevitably lead to a limited evaluation of context suitability. Both concepts leave open the question of whether the typicality judgment in a strong context, but adverse listening condition should be understood as a poorly generated or a more specific prediction. With our model, we suggest that these perspectives are two sides of the same coin.

To summarize, processing efforts in response to degraded auditory sentences capture resources that would normally be available for predictive processes in the mental lexicon in response to nondegraded sentences. Thus, we propose that adverse listening conditions limit the ability to form abstract expectancies from context, which leads to stronger reliance on acoustic-phonetic rather than lexical cues. This is in line with Mattys et al. (2009), who also demonstrated that, when listeners are confronted with energetically masked speech, they rely more on segmental (rather than lexical) information. Furthermore, studies on time-compressed speech (another form of speech degradation) have shown that listeners can recover intelligibility (i.e., access their mental lexicon) in severely time-compressed speech, as long as silent breaks are inserted at clause boundaries, such that listeners gain processing time intermittently (Wingfield, Tun, Koh, & Rosen, 1999). Bringing together this limited-resources account and the current results, an experimental prediction can be formed: By allowing the listener to free-up resources by allowing more time for processing the degraded sentence, the typicality-only effect in intermittent-delay degraded speech should be reduced.

To conclude, we propose a simplified, yet testable expectancy searchlight model (Figure 4), which aims to bring together the different aspects discussed here. While inevitably leaving open many questions (e.g., no assumption is formulated on how pre-lexical, acoustic-phonetic processing information enters the post-lexical stage), such a searchlight model is able to capture how expectancy is modulated by semantics and acoustics. It thereby combines top–down and bottom–up approaches.

Conclusions

This study investigated the relative importance of different sources of information in speech comprehension under adverse listening conditions. Do we rely more on top–down context or on bottom–up perceptual input? The data show that semantic context plays a crucial role, but deficient perceptual evidence in a degraded signal leads to more conservative, more narrowly adjusted expectancies on the forthcoming acoustic–phonetic information. Only common sentence endings are facilitated in the processing of moderately degraded speech. These results, thus, provide a starting point to better understand and aid speech comprehension in hearing-impaired and aging listeners.

REFERENCES

Aydelott, J., Baer-Henney, D., Trzaskowski, M., Leech, R., & Dick, F. (2012). Sentence comprehension in competing speech: Dichotic sentence-word priming reveals hemispheric differences in auditory semantic processing. *Language and Cognitive Processes, 27,* 1108–1144.

Aydelott, J., Dick, F., & Mills, D. L. (2006). Effects of acoustic distortion and semantic context on event-related potentials to spoken words. *Psychophysiology, 43,* 454–464.

Benki, J. R. (2003). Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition. *The Journal of the Acoustical Society of America, 113,* 1689–1705.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research, 27,* 32–48.

Boulenger, V., Hoen, M., Jacquier, C., & Meunier, F. (2011). Interplay between acoustic/phonetic and semantic processes during spoken sentence comprehension: An ERP study. *Brain and Language, 116,* 51–63.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58,* 412–424.

Cleland, A. A., Gaskell, M. G., Quinlan, P. T., & Tamminen, J. (2006). Frequency effects in spoken and visual word recognition: Evidence from dual-task methodologies. *Journal of Experimental Psychology: Human Perception and Performance, 32,* 104–119.

Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience, 6,* 256–266.

Connolly, J. F., Phillips, N. A., Stewart, S. H., & Brake, W. G. (1992). Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. *Brain and Language, 43,* 1–18.

Connolly, J. F., Stewart, S. H., & Phillips, N. A. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language, 39,* 302–318.

D'Arcy, R. C. N., Connolly, J. F., Service, E., Hawco, C. S., & Houlihan, M. E. (2004). Separating phonological and semantic processing in auditory sentence processing: A high-resolution event-related brain potential study. *Human Brain Mapping, 22,* 40–51.

D'Arcy, R. C. N., Service, E., Connolly, J. F., & Hawco, C. S. (2005). The influence of increased working memory load on semantic neural systems: A high-resolution event-related brain potential study. *Cognitive Brain Research, 22,* 177–191.

Davis, M. H., Ford, M. A., Kherif, F., & Johnsrude, I. S. (2011). Does semantic context benefit speech understanding through "top–down" processes? Evidence from time-resolved sparse fMRI. *Journal of Cognitive Neuroscience, 23,* 3914–3932.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience, 23,* 3423–3431.

Desroches, A. S., Newman, R. L., & Joanisse, M. F. (2009). Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity. *Journal of Cognitive Neuroscience, 21,* 1893–1906.

Eisner, F., McGettigan, C., Faulkner, A., Rosen, S., & Scott, S. K. (2010). Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. *The Journal of Neuroscience, 30,* 7179–7186.

Federmeier, K. D., & Kutas, M. (1999). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research, 8,* 373–392.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Méot, A., et al. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: Evidence from chronolex. *Frontiers in Psychology, 2,* 306. doi:10.3389/fpsyg.2011.00306.

Friedrich, C. K., & Kotz, S. A. (2007). Event-related potential evidence of form and meaning coding during online speech recognition. *Journal of Cognitive Neuroscience, 19,* 594–604.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology, 22,* 615–621.

Geyken, A. (2011). Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In A. Abel & R. Zanin (Eds.), *Korpora in Lehre und Forschung* (pp. 129–154), Bozen-Bolzano University Press.

Gunter, T. C., Friederici, A. D., & Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *Journal of Cognitive Neuroscience, 12,* 556–568.

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., et al. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *Neuroimage, 17,* 1101–1116.

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology, 30,* 47–61.

Howes, D. (1957). On the relation between the probability of a word as an association and in general linguistic usage. *Journal of Abnormal Psychology, 54,* 75–85.

Joordens, S., & Besner, D. (1992). Priming effects that span an intervening unrelated word: Implications for models of memory representation and retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 483–491.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–95). Cambridge, MA: MIT Press.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61,* 1337–1351.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4,* 463–470.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307,* 161–163.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience, 9,* 920–933.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics—Volume 2* (pp. 768–774). Stroudsburg, PA, USA. Association for Computational Linguistics. doi:10.3115/980432.980696.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19,* 1–36.

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology, 59,* 203–243.

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia, 50,* 762–776.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41,* 329–335.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2,* 217–228.

Newman, R. L., & Connolly, J. F. (2009). Electrophysiological markers of pre-lexical speech processing: Evidence for bottom–up and top–down effects on spoken word processing. *Biological Psychology, 80,* 114–121.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115,* 357–395.

Obleser, J., Eisner, F., & Kotz, S. A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. *The Journal of Neuroscience, 28,* 8116–8123.

Obleser, J., & Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex, 20,* 633–640.

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage, 55,* 713–723.

Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex, 22,* 2466–2477.

Obleser, J., Wise, R. J. S., Alex Dresner, M., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *The Journal of Neuroscience, 27,* 2283–2289.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 32,* 12376–12383.

Paulmann, S., Ott, D. V. M., & Kotz, S. A. (2011). Emotional speech perception unfolding in time: The role of the basal ganglia. *PloS One, 6,* e17694. doi:10.1371/journal.pone.0017694

Pichora-Fuller, M. K. (2003). Cognitive aging and auditory information processing. *International Journal of Audiology, 42(Suppl. 2),* 2S26–2S32.

Piquado, T., Cousins, K. A. Q., Wingfield, A., & Miller, P. (2010). Effects of degraded sensory input on memory for speech: Behavioral data and a test of biologically constrained computational models. *Brain Research, 1365,* 48–65.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology, 20,* 241–248.

Riecke, L., Esposito, F., Bonte, M., & Formisano, E. (2009). Hearing illusory sounds in noise: The timing of sensory-perceptual transformations in auditory cortex. *Neuron, 64,* 550–561.

Romei, L., Wambacq, I. J. A., Besing, J., Koehnke, J., & Jerger, J. (2011). Neural indices of spoken word processing in background multi-talker babble. *International Journal of Audiology, 50,* 321–333.

Schroeger, E., Trevaniemi, M., & Huotilainen, M. (2003). Bottom–up and top–down flows of information within auditory memory: Electrophysiological evidence. In C. Kaernbach, H. Mueller, & E. Schroeger (Eds.), *Psychophysics beyond sensation: Laws and invariants of human cognition* (pp. 389–410). Mahwah, NJ: Erlbaum.

Sivonen, P., Maess, B., & Friederici, A. D. (2006). Semantic retrieval of spoken words with an obliterated initial phoneme in a sentence context. *Neuroscience Letters, 408,* 220–225.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology: General, 112,* 1–36.

Stickney, G. S., & Assmann, P. F. (2001). Acoustic and linguistic factors in the perception of bandpass-filtered speech. *The Journal of the Acoustical Society of America, 109,* 1157–1165.

Tallon-Baudry, C., & Bertrand, O. (1999). Oscillatory gamma activity in humans and its role in object representation. *Trends in Cognitive Sciences, 3,* 151–162.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30,* 415–433.

van den Brink, D., Brown, C. M., & Hagoort, P. (2006). The cascaded nature of lexical selection and integration in auditory sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 364–372.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition, 18,* 380–393.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language, 97,* 279–293.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83,* 176–190.

Wingfield, A., Tun, P. A., Koh, C. K., & Rosen, M. J. (1999). Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech. *Psychology and Aging, 14,* 380–389.