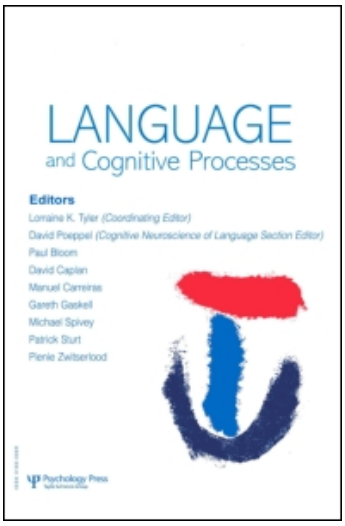

Language and Cognitive Processes
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713683153

Expectations in counterfactual and theory of mind reasoning

Heather J. Ferguson [a]; Christoph Scheepers [b]; Anthony J. Sanford [b]
[a] University College London, London, UK [b] University of Glasgow, Glasgow, UK

First Published on: 29 July 2009

To cite this Article Ferguson, Heather J., Scheepers, Christoph and Sanford, Anthony J.(2009)'Expectations in counterfactual and theory of mind reasoning',Language and Cognitive Processes,99999:1,

To link to this Article: DOI: 10.1080/01690960903041174

URL: http://dx.doi.org/10.1080/01690960903041174

LANGUAGE AND COGNITIVE PROCESSES
0000, 00 (00), 1–50

Psychology Press
Taylor & Francis Group

Expectations in counterfactual and theory of mind reasoning

Heather J. Ferguson
University College London, London, UK

Christoph Scheepers and Anthony J. Sanford
University of Glasgow, Glasgow, UK

During language comprehension, information about the world is exchanged and processed. Two essential ingredients of everyday cognition that are employed during language comprehension are the ability to reason counterfactually, and the ability to understand and predict other peoples' behaviour by attributing independent mental states to them (theory of mind). We report two visual-world studies investigating the extent to which the constraints of world knowledge and prior context, as established by a counterfactual (Exp. 1) or a false belief situation (Exp. 2), influence eye-movements directed towards objects in a visual field. Proportions of anticipatory eye-movements indicated an initial visual bias towards *contextually supported* referents in both studies. Thus, we propose that when visual information is available to reinforce linguistic input, participants *expect* a context-relevant continuation. Shortly after the critical word onset, the *linguistically supported* referent was visually favoured, with counterfactual (but not false belief) contexts revealing a temporal delay in integrating factually inconsistent language input. Results are discussed in relation to accounts of discourse processing and the processing relationship between counterfactual and theory of mind reasoning. Finally, we compare findings across different experimental paradigms and propose a novel cluster-analytic procedure to identify time-windows of interest in visual-world data.

Keywords: Counterfactual reasoning; Discourse processing; Theory of mind; Visual-world paradigm; *k*-means clustering.

Correspondence should be addressed to Heather Ferguson, UCL Research Department of Linguistics, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK. E-mail: h.ferguson@ucl.ac.uk

The order of the first two authors is arbitrary. HF acknowledges the support of ESRC in the form of a postgraduate studentship. Thanks are due to members of the Glasgow Language Group, Gerry Altmann and three anonymous reviewers for helpful comments on this work.

http://www.psypress.com/lcp DOI: 10.1080/01690960903041174

The ability to update our current knowledge using contextual information is a vital process during everyday language comprehension. We seem immediately able to use relevant linguistic and non-linguistic information, such as the wider discourse and its genre, as well as the intentions, beliefs and desires of others, to enhance comprehension of an unfolding sentence (e.g., Van Berkum, 2008; Van Berkum, Holleman, Murre, Nieuwland, & Otten, 2007; Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008). Counterfactual and theory of mind reasoning are two essential ingredients of our everyday cognition, yet surprisingly little is known of how they are processed on-line during reading or listening. Investigations of counterfactuals and theory of mind are particularly timely in light of a recent debate in the literature on children's reasoning that considers the extent to which counterfactuals and false beliefs share cognitive processes (Perner, Sprung, & Steinkogler, 2004; Peterson & Bowler, 2000; Riggs, Peterson, Robinson, & Mitchell, 1998). In this paper, we attempt an exploration of processing during the comprehension of counterfactuals and the beliefs of others in a visual world task.

Counterfactuals are cases of possibly valid reasoning from premises that are false in actuality (Fauconnier & Turner, 2003), and require the comparison of reality to a model-based alternative. People understand a counterfactual statement, such as in *If Lucy had worked hard she would have passed her exams*, by keeping in mind two possibilities from the outset: the conjecture, *Lucy worked hard* and *she passed her exams*, and the presupposed facts, *Lucy did not work hard* and *she did not pass her exams* (Byrne & Tasso, 1999). The counterfactual thus requires that a person represent false information that is temporarily supposed to be true. There is a considerable literature available on reasoning with counterfactuals (see Byrne, 2002), and on what sort of constraints there are on the kinds of counterfactual thoughts people are likely to generate in a variety of circumstances (e.g., Byrne, 1997; Kahneman & Miller, 1986; Markman & Tetlock, 2000). However, within the framework of reasoning and its social concomitants there has been limited research on how counterfactuals are processed on-line during language comprehension.

Recently, Ferguson and Sanford (2008); see also Ferguson, Sanford, & Leuthold, 2006) have used a novel approach to examine this issue using eye-movement investigations in reading. Participants read short passages where a context sentence introduced a counterfactual-world (CW), as in *If cats were vegetarians* . . ., or a real-world (RW) situation, as in *If cats are hungry* . . .; then a second sentence was manipulated to create RW-incongruent/CW-congruent continuations (e.g., *Families could feed their cat a bowl of carrots* . . .), respectively RW-congruent/CW-incongruent continuations (e.g., *Families could feed their cat a bowl of fish* . . .). Results showed that typical effects of real-world violations can be 'neutralised' within an appropriate pre-specified CW context. Further, RW-congruent items can

lead to the experience of an anomaly following a CW context. Importantly, there was also evidence for early processing difficulty with RW violations regardless of prior context, indicating that a proposition is rapidly evaluated against real-world knowledge, just prior to the accommodation of a proposition into a counterfactual world representation. These results support a dual, possibly two-stage, comprehension process for counterfactuals, where both factual and counterfactual information is available to the reader. Clearly, when the use of a word violates real-world knowledge, this creates a very early effect upon reading, while contextual information appears to influence later discourse resolution.

The above findings are compatible with the mental model theory of Johnson-Laird (1983); Johnson-Laird & Byrne, 1991), which assumes that different mental spaces are created to represent information during language comprehension. This theory has a 'core' extensional account of conditionals, making a conditional *if p then q* logically equivalent to *not-p or q*. Thus, mental spaces, and the relationships between them, are a way of specifying an interpretation of a discourse. The mental model theory has been applied to counterfactual reasoning (Fauconnier, 1985, 1997). According to Fauconnier, two mental spaces are produced in the case of counterfactual conditionals; one is the reality space and the other is the counterfactual hypothetical space. Hence, counterfactuality is described as a case of forced incompatibility between these two spaces, since what is true in the counterfactual space is false in the reality space.

Similar to counterfactual reasoning, theory of mind is a case of possibly valid reasoning based on the beliefs of other people that might be false according to our own knowledge of reality. Tasks involving theory of mind (ToM) require an understanding of events according to the intentions, beliefs, and desires of other people. Much work on ToM has centred around impairments of this ability, such as autism spectrum disorders (Baron-Cohen, 2000; Leslie, 1994; Tager-Flusberg, Boshart, & Baron-Cohen, 1998) and schizophrenia (Frith & Corcoran, 1996; Frith & Frith, 1988) and also on locating a neurological basis for ToM reasoning (Gallagher & Frith, 2003; Happé, Malhi, & Checkley, 2001; Rowe, Bullock, Polkey, & Morris 2001; Stone, Baron-Cohen, & Knight, 1998; Stuss, Gallup, & Alexander, 2001). However, this research has been limited by the use of traditional response-based measures used to investigate ToM comprehension.

Since ToM situations require comprehenders to represent information about both their own reality and reality according to another person's beliefs, it appears plausible to assume that ToM tasks engage a dual comprehension process involving multiple mental spaces, similar to counterfactuals. In fact, such a link between ToM and counterfactuals has recently been proposed by developmental theorists who suggest that theory of mind is a special case of counterfactual thinking, and as such may engage a network of consistent

specialised cognitive processes (Leslie, 1987; Riggs et al., 1998). Indeed, ability in counterfactual reasoning has emerged as a necessary but not sufficient component of successful performance in false belief tasks (Peterson & Bowler, 2000). However, although this theoretical similarity appears plausible given the multiple mental spaces required by both counterfactuals and ToM, this is a relatively novel link and as such, investigations of it have been very limited. Specifically, no empirical studies have examined the nature of cognitive processes involved in creating, maintaining, and selecting the appropriate representation in each case. The present paper is a direct attempt to examine the role played by real-world (factual) knowledge, and inferences from counterfactual worlds (Exp. 1) or the beliefs of others (Exp. 2) during on-line comprehension of simple statements in a 'visual world' setting, and to determine whether related processes are revealed in these tasks.

Recently, some studies have attempted to use on-line measures to investigate the linguistic processing of perspectives, including common ground knowledge and ToM issues (Epley, Morewedge, & Keysar, 2004; Hanna, Tanenhaus, & Trueswell, 2003; Keysar, Barr, Balin, & Brauner, 2000; Keysar, Lin, & Barr, 2003; Nadig & Sedivy, 2002). Common ground refers to knowledge shared by two or more interlocutors and differs from privileged ground knowledge, which represents knowledge known to only one member of a group. These studies employed a version of the visual-world paradigm: participants' eye-movements were monitored while they followed a confederate's instructions to manipulate real-world objects. Using this method, evidence has been provided that communicators have rapid access to common ground information and can use perspective cues to accurately infer privileged information from a speaker. Keysar et al. (2003) used this technique to enhance current understanding of ToM, reporting a dissociation between peoples' ability to reflect on information from their own versus other peoples' knowledge and the routine ability to apply it in social situations. Their results support an egocentric view of ToM processing by suggesting that while people have no problem assessing another person's knowledge, doing so is cognitively costly and thus conversation is frequently grounded in information from one's own perspective. Thus, Keysar et al. propose that communicators do not consistently use information about others' knowledge, intentions, or desires to predict the actions of that person. Further, recent research suggests that speakers' choices of syntactic structure are often made without consideration of listeners' needs (Arnold, Wasow, Asudeh, & Alrenga, 2004; Ferreira & Dell, 2000; but cf. Haywood, Pickering, & Branigan, 2005). However, only one study to date has examined the progressive temporal nature of perspective switches (Hanna et al., 2003). Using an improved version of the Keysar et al. (2003) design, which limited confounds from recency and likelihood of mention, Hanna and colleagues have provided preliminary evidence for simultaneous integration of perspectives

during comprehension. Thus, they propose a constraint-based view where, despite interference from the privileged view, communicators can immediately predict reference to a common ground competitor. Importantly, this study focuses on integration of information based on perspectives but does not directly test whether knowledge of another person's beliefs can lead to assumptions on predicting others' behaviour. These issues will be addressed in the current paper.

The present studies investigate the extent to which the constraints of real-world knowledge and prior context influence eye-movements directed towards entities in a visual field. In recent years large amounts of research have used the visual-world paradigm to investigate language-mediated eye-movements. With such a technique, initial studies demonstrated that eye-movements can be directed by auditory input towards appropriate objects in a visual display (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). It is commonly believed that such language-mediated eye-movements reflect the cognitive processes that underlie language comprehension. Moreover, growing evidence has shown that the eyes can move towards a critical object *before* a word referring to that object is available. Altmann and Kamide (1999), for example, demonstrated that when an auditory sentence such as '*the boy will eat the cake*' is paired with a visual display depicting a boy, a cake, and some inedible objects, listeners launch anticipatory eye-movements towards the cake as soon as the verb '*eat*' is available.

Numerous experiments have used the visual world paradigm to demonstrate that discourse processing is driven by predictive relationships involving syntax (e.g., Arai, van Gompel, & Scheepers, 2007; Scheepers & Crocker, 2004; Kamide, Scheepers, & Altmann, 2003), semantics (e.g., Huettig & Altmann, 2005; Kamide, Altmann, & Haywood, 2003; Scheepers, Keller, & Lapata, 2008; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Yee & Sedivy, 2006) and real-world expectations (Altmann & Kamide, 2009). However, it is not clear whether these fundamental processing strategies can also be influenced by introducing a counterfactual context or by manipulating the beliefs of others (ToM). The visual-world paradigm is ideal to study top-down expectations in these two types of reasoning, as both counterfactuals and the beliefs of others may rapidly establish a context that is contradictory to the readers' real-world or factual knowledge. As such, counterfactuals and knowledge of the beliefs of others may create strong predictive biases that would interact with real-world knowledge during processing. The issue we examined, therefore, was whether people can use their knowledge of the wider discourse to override real-world knowledge to predict specific upcoming reference as the current sentence is unfolding. Thus, the current experiments complement the work from Ferguson and Sanford (2008) that

specifically looked at integration and recovery strategies following counterfactually licensed pragmatic anomalies.

In sum, the studies reported here offer a comparison of counterfactual conditional reasoning with reasoning based on the beliefs of others (ToM). The motivation for this design originates from the potential commonalities and differences between counterfactual and ToM contexts. Specifically, both tasks require comprehenders to represent two versions of the world; one is factual or reality-based and the other is derived from some counterfactual scenario or privileged information about the beliefs of others. Further, successful interpretation of both counterfactual and ToM contexts involves ignoring 'what you know' to focus on an alternative state of affairs, thus demonstrating that deductive reasoning plays an integral role in their comprehension. In contrast, a critical difference between counterfactual and ToM reasoning tasks is the fact that ToM tasks implicate characters' mental states. Counterfactuals, on the other hand, typically focus on physical states. Therefore, it is likely that the mental state element involved in ToM tasks requires extra cognitive processes compared to counterfactual reasoning. The current experiments will attempt to disentangle these effects and explore how counterfactual and false belief contexts influence expectations that are derived online during language comprehension.

EXPERIMENT 1

Experiment 1 investigated the comprehension of counterfactual conditionals such as in (1).

(1) If cats were vegetarians ...

Participants heard a real-world (RW) or counterfactual-world (CW) context sentence, followed by a target sentence that was paired with visually presented referents. Eye-movements around the visual scene were monitored and time-locked to related auditory input to examine context effects on the anticipation of forthcoming linguistic RW or CW referents. According to the mental model theory, counterfactual reasoning requires people to keep in mind both the counterfactual and the factual alternatives. This immediately leads to a processing question: can our real-world expectations be 'neutralised' within a pre-specified counterfactual world context so that comprehenders immediately predict upcoming linguistic input according to the preceding CW context? Further, if this context- bound prediction does emerge, at what stage of processing is it revealed? More specifically, can a prior counterfactual context lead to anticipatory eye-movements towards contextually relevant objects in a scene (that are anomalous given RW knowledge) or is this

contextual integration process delayed so that it initially leads to a RW preference, and later becomes accommodated by the counterfactual world representation? This is the basic question of Experiment 1.

Method

Participants. Twenty-eight participants from the University of Glasgow's undergraduate population were paid to take part in the study. All were native English speakers with normal or corrected to normal vision and had no prior exposure to the experimental items. The same 28 participants also took part in Experiment 2. Note that the two experiments were run in separate testing blocks in a counterbalanced order (half of the participants received Experiment 1 first and the other half Experiment 2 first), alongside different filler items.

Stimuli and design. Twenty-four experimental pictures were paired with auditory passages in one of four conditions. Table 1 and Figure 1 provide an example of such experimental sentences and the associated visual displays. The latter were created using commercially available clip-art collections and were presented on a 21-inch colour monitor running at 85 Hz refresh rate in 1024×768 pixels resolution. Each scene contained four objects: Topic (the cat in the given example), RW Referent (fish), CW Referent (carrots), and a Distracter (bus) which was neither RW nor CW congruent. To prevent any systematic viewing strategies, spatial arrangements of these four picture elements differed across items. Sound files consisted of two sentences: Sentence one created a RW or CW context (*If cats are hungry* ... versus *If cats were vegetarians* ...) and Sentence two drew reference to a RW- or CW-consistent referent (*Families could feed their cat a bowl of fish* versus *carrots* ...), resulting in a 2×2 within subjects design. Importantly, all items used concepts paired with highly predictable associates (e.g., cats–fish / vegetarians–carrots; America–baseball / Spain–bullfights; spider–web / bee–honey) to narrow the number of alternatives available to perceivers.[1] Note that CW-consistent referents (e.g., *carrots*) were anomalous in RW contexts, and vice-versa for RW-consistent referents (e.g., *fish*). Experimental sentences varied in syntactic structure, such that the critical word ('fish' or 'carrots') did not always occur in exactly the same position across items. However, we made sure that the position of the critical word always occurred

[1] The predictability of these associates is supported by Ferguson and Sanford's (2008) eye tracking studies. In an additional pre-test, we collected word association ratings ranging from 1 (low word association) to 5 (high word association). The resulting mean scores were 4.6 (RW-RW), 1.2 (RW-CW), 4.1 (CW-CW) and 1.1 (CW-RW), with no significant differences between RW-RW and CW-CW scores or RW-CW and CW-RW scores (all $ts < 1$), suggesting high association for concept matched words and low association for concept mismatched words.

TABLE 1
Examples of experimental sentences (Experiment 1)

RW context – CW language input
If cats are hungry they usually pester their owners until they get fed.
Families could feed their cat a bowl of carrots and it would gobble it down happily.
RW context – RW language input
If cats are hungry they usually pester their owners until they get fed.
Families could feed their cat a bowl of fish and it would gobble it down happily.
CW context – CW language input
If cats were vegetarians they would be cheaper for owners to look after.
Families could feed their cat a bowl of carrots and it would gobble it down happily.
CW context – RW language input
If cats were vegetarians they would be cheaper for owners to look after.
Families could feed their cat a bowl of fish and it would gobble it down happily.

roughly mid-sentence and was identical across conditions for each item. One version of each item was assigned to one of four presentation lists, with each list containing 24 experimental items, 6 in each of the four conditions, blocked to ensure that they were evenly distributed. In addition, 24 unrelated filler items were added to each list.[2] They all consisted of correctly matched picture-sentence pairings and were interspersed randomly among the 24 experimental trials to create a single random order. Each subject only saw each target sentence once, in one of the four conditions. At least one filler trial intervened between any two experimental trials.

Sentences were recorded in one session from a female native British English speaker who was instructed to use a neutral intonation. The auditory files were presented as 16 KHz mono sound clips via a satellite speaker system connected to the eye-tracker PC. The temporal onsets of critical words in Sentence 2 were hand-coded with millisecond resolution using the GoldWave sound-editing package.

Comprehension questions, relating to either the auditory or visual input, followed half of the experimental and half of the filler trials. Participants did not receive feedback for their responses to these questions. All participants scored at or above 90% accuracy on the comprehension questions.

Procedure. Participants were seated in front of a 21-inch colour monitor that was connected to an SR Research Eyelink II head-mounted eye-tracking system running at 500 Hz sampling rate. Viewing was binocular, but only the participants' dominant eye was tracked, as determined via a simple parallax test prior to the experiment. Participants were given the following

[2] The full list of filler items can be obtained from the first author.

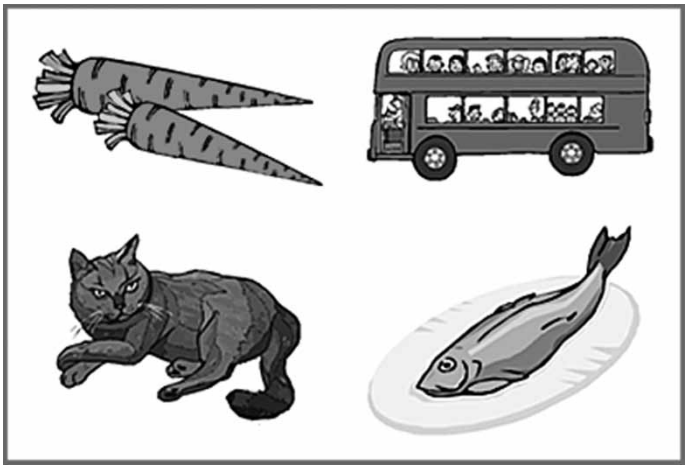

Figure 1. Example visual stimulus used in Experiment 1. Participants heard the target sentence (see above) whilst viewing this picture.

instruction: 'In this experiment you will hear short spoken passages and during the second sentence a picture will also be displayed. We are interested in how the pictures help you understand the spoken passages'.

As illustrated in Figure 2, each trial began with the presentation of a single centrally located dot and participants were asked to fixate it so that an automatic drift correction could be performed. While the participant fixated this dot, the experimenter pressed a button to initiate the trial. The dot was replaced by a fixation cross while participants heard Sentence 1 (RW or CW

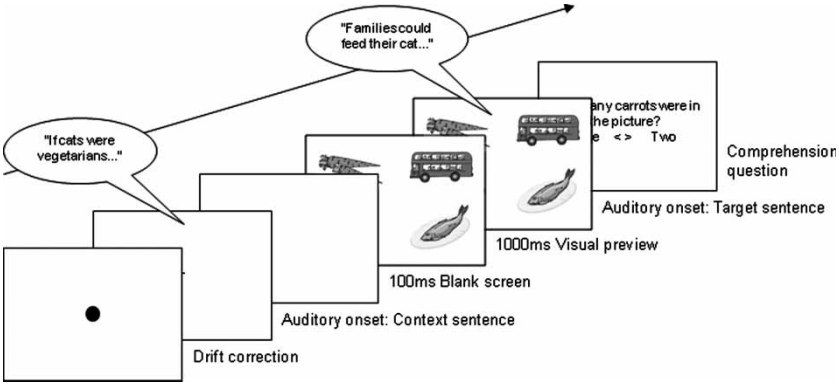

Figure 2. Illustration of the experimental procedure in Experiment 1.

context, no picture presentation). They were asked to continue looking at the fixation cross during this time. Then a 100 ms blank screen was presented, followed by the target picture combined with Sentence 2. The onset of the target picture preceded the onset of the corresponding spoken sentence by 1000 ms. The picture stayed on the screen for 9 s, and the corresponding sentence typically ended 1–2 s before the end of the trial.

At the beginning of the experiment, and once every ten trials thereafter, the eye-tracker was calibrated and validated against nine fixation points. This procedure took about half a minute and an entire session lasted for about half an hour.

Results and discussion

Data processing. Eye-movements that were initiated during Sentence 2 were processed according to the relevant picture and sound onsets for the purpose of aggregating fixation locations and durations. Temporal onsets and offsets of the fixations were recalculated relative to the corresponding picture onset by subtracting the picture onset from the relative fixation onsets. An automatic procedure was used to pool short contiguous fixations. Fixations shorter than 80 ms (fewer than 4% of the cases) were integrated with the immediately preceding or following fixation if that fixation lay within half a degree of visual angle, otherwise the fixation was excluded. The rationale for this was that such short fixations usually result from false saccade planning (see Rayner & Pollatsek, 1989) and are unlikely to reflect meaningful information processing. In case a blink occurred, its duration was added to the immediately preceding fixation (processing is unlikely to pause during a blink). The spatial coordinates of the fixations (in pixels) were then mapped onto the appropriate object regions using colour-coded bitmap templates; if a fixation was located within 20 pixels around an object's perimeter, it was coded as belonging to that object, otherwise, it was coded as background. Finally, all consecutive fixations within one object region before the eyes moved to a different region were pooled into a single *gaze*.

As with Arai et al. (2007), we analysed probabilities of gazes to the critical RW and CW referents as a function of time, using the following log-ratio measure:

$$\log(\mathrm{RW}/\mathrm{CW}) = ln(P_{(\mathrm{RW})}/P_{(\mathrm{CW})}), \quad (1)$$

where $P_{(\mathrm{RW})}$ refers to the probability of gazes on the RW referent (fish) and $P_{(\mathrm{CW})}$ to the probability of gazes on the CW referent (carrots); *ln* refers to the natural logarithm. The measure is symmetrical around zero such that higher proportions of gazes on the RW referent result in a positive score, e.g., $ln(.50/.25) = 0.693$, and higher proportions of gazes on the CW referent in a

negative score, e.g., $ln(.25/.50) = -0.693$. Equal proportions of looks between the two referents yields a score of zero, $ln(1) = 0$.[3]

For our log(RW/CW) analyses, we chose a time period ranging from 1500 ms *before* the onset of the critical word ('fish' or 'carrots', respectively) until 2000 ms *after* the onset of the critical word. Across items, 95% CIs for the critical word period (measured from the onset of the critical word to the onset of the subsequent word) amounted to 931 ± 31 ms and 919 ± 37 ms for the RW and CW language input conditions, respectively. The data were synchronised on a by-trial basis, relative to the onset of the critical word in the appropriate item-condition combination. Figure 3 plots the observed average log(RW/CW) data in each condition, for every 20 ms time-slot within the selected time period. The solid black vertical line in the figure ($t = 0$) indicates the critical word onset and the arrow indicates the average verb onset (e.g., *feed*).

In order to reduce the number of statistical tests without masking potentially important detail, the 20 ms time-slots in Figure 3 were aggregated into larger analysis windows. We employed a cluster-analytic procedure (see Appendix for full details of this procedure) which identified nine groups of contiguous time-slots (among the 175 available) that showed maximally similar cross-condition data patterns within each cluster and maximally distinct patterns between clusters. Since one of these clusters spanned across the critical word onset (-200 to 400 ms), we divided it into two clusters for analysis. This ensured a clear-cut distinction between effects immediately preceding (i.e., anticipation) and following the critical word onset. The resulting analysis clusters are indicated by the dashed vertical lines in Figure 3.

Note that with the current dataset, typical word-based analyses (cf., Altmann & Kamide, 1999) were not deemed appropriate as the key manipulation here concerned the context *prior* to the critical sentence rather than verb-bound semantic constraints within the sentence. As such, potential effects of our manipulation were not expected to be tightly bound to individual 'triggering' constituents in the target sentence. Further, as Figure A (see Appendix) demonstrates, the word-based approach is of limited power for visual-bias transitions *within* the critical word-region.

Main analyses. In the first set of analyses, we were interested in whether prior context (RW versus CW) affected proportions of gazes on the RW referent (the *fish* in Figure 1) relative to the CW referent (the *carrots* in Figure 1) in time periods preceding and following the onset of the critical

[3] Since this measure only takes proportions of gazes on the RW and CW referent into account, it is important to stress that proportions of gazes on the Topic (cat) and Distracter (bus) referents revealed no significant cross-condition effects whatsoever. See http://www.phon.ucl.ac.uk/research/ferguson.pdf for the corresponding raw probability plots (separately for critical and non-critical objects).

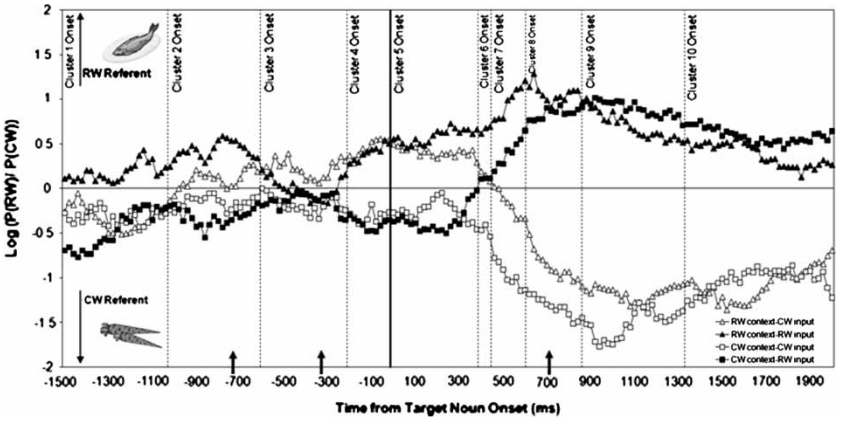

Figure 3. The average log(RW/CW) as a function of each condition in Experiment 1. Note that the solid black vertical line in the figure (t = 0) indicates the target noun onset, dashed lines represent cluster boundaries for statistical analysis and the arrows indicate (from left to right) the average verb onset and offset (e.g., feed), and the average target noun offset.

word ('fish' or 'carrots'). Another question was whether and how type of context interacted with the RW/CW language input in Sentence 2.

For each participant (and respectively item) and condition, a weighted average[4] log(RW/CW) score was calculated over the 20 ms time slots per analysis cluster (dashed vertical lines in Figure 3). The weighted averages per cluster were then subjected to 2×2 ANOVAs with context (RW vs. CW) and language input (RW vs. CW) as repeated-measures factors. Table 2 displays the statistical details of the effects, allowing generalisation to participants (F_1) and items (F_2), for each time window of interest. Strength of association is reported in terms of partial eta-squared ($_p\eta^2$).

The analyses revealed no fully consistent effects within the first three time windows (cluster 1 to 3). However, a reliable main effect of context emerged in cluster 4, beginning 200 ms prior to the critical word onset and ending right at the critical word onset.[5] Fixations were more likely to be made towards contextually relevant referents. That is, a RW context led to an anticipatory visual bias towards the RW-referent (as indicated in more positive log(RW/CW) scores in Figure 3) and a CW context lead to an anticipatory bias towards the CW-referent (negative log(RW/CW) scores).

[4] Due to saccades or occasional blinks, numbers of observations differed slightly across 20 ms time slots. The weighted average takes this into account, so that time slots with more observations contribute proportionally more to the average than time slots with fewer observations.

[5] Note that this cluster roughly relates to the post-verb region using a word-based analysis approach. See Table A in the Appendix for full statistical details using this approach.

TABLE 2
Analysis of variance results for each time window of interest (Experiment 1)

	F_1				F_2			
Source of variance	*df*	F_1 *value*	*p-value*	$_p\eta^2$	*df*	F_2 *value*	*p-value*	$_p\eta^2$
(−1500)–(−1020) ms								
Context	1, 27	1.31	.26	0.05	1, 23	6.28	.02*	0.21
Language input	1, 27	1.7	.2	0.06	1, 23	2.66	.12	0.1
Context × Language input	1, 27	0.91	.35	0.03	1, 23	1.7	.21	0.07
(−1020)–(−600) ms								
Context	1, 27	4.2	.05*	0.14	1, 23	1.46	.24	0.06
Language input	1, 27	1.95	.17	0.07	1, 23	0.12	.73	0.01
Context × Language input	1, 27	1.64	.21	0.06	1, 23	0.87	.36	0.04
(−600)–(−200) ms								
Context	1, 27	6.8	.01**	0.19	1, 23	2.41	.13	0.1
Language input	1, 27	0.68	.42	0.03	1, 23	0.12	.73	0.01
Context × Language input	1, 27	0.75	.39	0.03	1, 23	0	.99	0
(−200)–0 ms								
Context	1, 27	21.02	<.001***	0.44	1, 23	6.18	.02*	0.21
Language input	1, 27	0.08	.78	0	1, 23	0.11	.74	0.01
Context × Language input	1, 27	0.14	.72	0.01	1, 23	0.91	.35	0.04
0–400 ms								
Context	1, 27	5.9	.02*	0.18	1, 23	13.32	<.001***	0.37
Language input	1, 27	0.21	.65	0.01	1, 23	0.37	.55	0.02
Context × Language input	1, 27	0.34	.56	0.01	1, 23	0.67	.42	0.03

Continued

TABLE 2 (*Continued*)

	F_1				F_2			
Source of variance	*df*	*F_1 value*	*p-value*	*$_p\eta^2$*	*df*	*F_2 value*	*p-value*	*$_p\eta^2$*
400–460 ms								
Context	1, 27	3.47	.07	0.11	1, 23	6.7	.02*	0.23
Language input	1, 27	1.51	.23	0.05	1, 23	1.73	.2	0.07
Context × Language input	1, 27	0.11	.75	0.01	1, 23	0.21	.65	0.01
460–620 ms								
Context	1, 27	7.38	.01**	0.22	1, 23	8.56	.008**	0.27
Language input	1, 27	16.55	<.001***	0.38	1, 23	31.13	<.001***	0.53
Context × Language input	1, 27	0.5	.48	0.02	1, 23	0.07	.79	0
620–880 ms								
Context	1, 27	1.77	.19	0.06	1, 23	2.63	.12	0.1
Language input	1, 27	45.64	<.001***	0.63	1, 23	67	<0.001***	0.74
Context × Language input	1, 27	1.06	.31	0.04	1, 23	0.06	.81	0
880–1340 ms								
Context	1, 27	0.02	.9	0	1, 23	0.03	.86	0
Language input	1, 27	45.6	<.001***	0.63	1, 23	55.43	<.001***	0.71
Context × Language input	1, 27	3.93	.06	0.13	1, 23	0.99	.33	0.04
1340–2000 ms								
Context	1, 27	0.91	.35	0.03	1, 23	0.5 9	.45	0.03
Language input	1, 27	16.6	<.001***	0.33	1, 23	40.9	<.001***	0.64
Context × Language input	1, 27	3.43	.07	0.11	1, 23	0.64	.43	0.03

Note: *$p < .05$; **$p < .01$; ***$p < .001$.

The main effect of language input was neither expected nor found during this time period where the critical word is not yet available. The effect of context persisted into cluster 5 (0–400 ms post-critical word onset) with the same pattern of gazes as in cluster 4. Again, no main effect of language input emerged in this time window, which suggests that the critical word is still being processed until up to 400 ms after critical word onset. Also note that an additional inspection of the *intercept* estimates for the ANOVAs up to and including cluster 5 (testing whether the log-ratio scores are generally different from zero) revealed no evidence for an overall visual bias (all Fs < 2). This suggests that the observed anticipation effects are *solely* driven by context (additional consideration of world knowledge should have manifested itself in reliably more positive log-ratio scores overall).

The context effect persisted into cluster 7 (460–620 ms post-critical word onset), as more fixations were made towards contextually relevant referents. However, during this time period, a main effect of language input also emerged, indicating that the critical word has been recognised and that the relevant 'appropriate' referent is therefore visually favoured. In other words, participants' attention has shifted from purely contextually constrained expectations to additional bottom-up influences from the available language input. Subsequent analyses revealed a similar main effect of language input in cluster 8 (620–880 ms), cluster 9 (880–1340 ms) and cluster 10 (1340–2000 ms). This suggests that from *c*. 460 ms after the onset of the critical word, participants visually favour the referent that is 'appropriate' to the auditory input, regardless of inconsistencies with prior context.

Interestingly, Figure 3 also suggests that after perceivers had recognised the critical word (from cluster 7 onwards), the resulting bias towards RW referents peaked earlier (between cluster 7 and 8, see positive scores) than the bias towards CW referents (peaking within cluster 9, see negative scores) – independently of context (there were no context × language input interactions for these clusters, cf. Table 2). Indeed, this could reflect an interesting difference in the average timing of responses to RW versus CW-consistent language input, specifically given the following considerations. Participants usually display considerable variability in how fast they respond to a stimulus (such as a critical word in a sentence). Consequently, the average response is likely to be distributed over time in a shape that follows a peak distribution function of some sort, where the location of the peak would indicate the point in time where the majority of perceivers have responded. Alternatively, since we are dealing with continuously varying visual biases here, the peak location may indicate the point in time where the response is strongest in individual perceivers. Either way, some important meaning can be ascribed to the peak location, as it indicates the point in time of the 'bulk' of the response (in terms of numbers of observations or in terms of signal

strength or in terms of a combination of the two) and thus indicates critical timing differences between conditions.

To address this issue, we conducted additional analyses with language input (RW vs. CW, averaged across the two context conditions) and cluster (7, 8, 9, 10) as repeated-measures factors.

The analyses were performed on *cluster ranks* per language input condition, computed individually for each participant and item, respectively. Since four clusters were considered, the rank scores ranged from 1 to 4. For RW language input, the cluster with the most *positive* log-ratio mean was scored highest (4) and the cluster with the least positive log-ratio mean lowest (1); for CW language input, the cluster with the most *negative* log-ratio mean was scored highest (4) and the cluster with the least negative log-ratio mean lowest (1). In this way, the cluster ranks indicated (on an ordinal scale) how strongly perceivers were biased towards the 'appropriate' referent in each time-cluster after recognising the critical word.[6] The mean cluster ranks are shown in Table 3, together with 95% confidence limits both for the rank means and for the RW-CW rank differences per cluster.

Two-way ANOVAs established a main effect of cluster, $F_1(3, 81) = 9.72$, $p < .001$; ${}_p\eta^2 = 0.26$; $F_2(3, 69) = 4.12$, $p < .01$; ${}_p\eta^2 = 0.15$, due to differing average ranks across the four clusters. The main effect of language input was not, and could not be, statistically meaningful because the coding implied an average rank of exactly $(1+2+3+4)/4 = 2.5$ in each language input condition. Crucially, there was a reliable cluster × language input interaction, $F_1(3, 81) = 3.50$, $p < .02$; ${}_p\eta^2 = 0.11$; $F_2(3, 69) = 4.98$; $p < .004$; ${}_p\eta^2 = 0.18$; inspection of Table 3 indicates that early on (cluster 7 and 8) the ordinal bias towards 'appropriate' referents tended to be stronger with RW rather than CW language input; later (cluster 9 and 10) the reverse was true. Hence, this analysis confirms a reliably earlier peak location for the bias towards 'appropriate' referents after recognising RW- rather than CW-congruent language input. This is interesting because it represents an analogy to findings from reading showing that real-world consistent information is easier to integrate early on than real-world inconsistent information, regardless of context (Ferguson & Sanford, 2008).

Preview region. In order to examine very early anticipatory effects prior to language input, we analysed log(RW/CW) distributions during the picture preview. Recall that the onset of the picture preceded the onset of the

[6] The main advantage of using cluster ranks is that they allow for testing differences in peak location (temporal cluster in which the relevant visual bias reaches its maximum) independently of any differences in peak amplitude (overall 'strength' of visual bias). Moreover, rank scores are very robust against extreme values.

TABLE 3
Mean cluster ranks representing ordinal strength of bias towards 'appropriate' referents per cluster as a function of language input (RW vs. CW congruent). Ninety-five per cent confidence limits (by participants and items) are provided for the cluster ranks themselves as well as for the RW-CW input difference per cluster

	Cluster rank (by subjects)	*Cluster rank (by items)*
Cluster 7 (460–620 ms)		
RW input	2.07 ± 0.41	2.21 ± 0.47
CW input	1.61 ± 0.34	1.79 ± 0.48
Difference	+0.46 ± 0.54	+0.42 ± 0.51
Cluster 8 (620–880 ms)		
RW input	3.04 ± 0.36	3.21 ± 0.28
CW input	2.54 ± 0.39	2.46 ± 0.37
Difference	+0.50 ± 0.52	+0.75 ± 0.52
Cluster 9 (880–1340 ms)		
RW input	2.82 ± 0.37	2.58 ± 0.48
CW input	3.11 ± 0.30	2.92 ± 0.45
Difference	−0.29 ± 0.43	−0.33 ± 0.59
Cluster 10 (1340–2000 ms)		
RW input	2.07 ± 0.48	2.00 ± 0.50
CW input	2.75 ± 0.48	2.83 ± 0.46
Difference	−0.68 ± 0.68	−0.83 ± 0.67

corresponding target sentence by 1000 ms. Specifically, we analysed weighted average log(RW/CW) scores from 500 ms to 1000 ms post-picture onset (prior to that time period, participants were likely to fixate the area around the previously presented fixation cross, meaning insufficient numbers of observations for log(RW/CW) analyses). Note that this time window did not overlap with the main analysis time period in any of the experimental trials. The analyses revealed a main effect of context, $F_1(1, 27) = 4.2$, $p < .05$, ${}_p\eta^2 = 0.12$; $F_2(1, 23) = 9.2$, $p < .01$, ${}_p\eta^2 = 0.29$, which was mainly due to increased proportions of gazes on the CW-referent following a CW context (95% CIs for the log(RW/CW) scores in this condition: -0.36 ± 0.32 by subjects; -0.55 ± 0.49 by items), whereas following a RW context, no significant anticipation of the corresponding RW referent was evident (0.13 ± 0.32 by subjects; -0.04 ± 0.53 by items). This suggests that after a CW context, participants are already forming assumptions as to an appropriate continuation even before the target sentence is available. The main effect of language input was not reliable during the preview period (Fs < 1.2) and there was no significant context by language input interaction either (Fs < 0.5).

To summarise, participants were able to quickly use prior context information to make anticipatory eye-movements towards a relevant referent

in the visual display. The corresponding anticipation effects were initiated from at least 200 ms prior to critical word onset. Our data indicated that participants expect a context-relevant continuation *regardless* of whether this continuation is consistent with world-knowledge or not. Indeed, additional analyses of the preview region revealed that following a counterfactual context (but not following a real-world context) participants were immediately drawn towards the contextually relevant visual object. This suggests that the mental space representing counterfactual information expects a continuation of the form 'If ... then ...', whereas a real-world context does not seem to trigger immediate expectations towards a specific consequence. This difference could be due to the novelty of the counterfactual scenario set up by the CW context, which becomes more salient to comprehenders than conventional real-world scenarios.

At least for the given example, a potential concern might be that the word 'vegetarians' in the CW context conditions could prime access to *carrots* in the visual scene. However, the following considerations render low-level priming a rather implausible explanation of the anticipatory eye-movement patterns during the preview period. First, the CW context not only mentioned 'vegetarians' (related to *carrots*) but also 'cats' (related to *fish*). Second, the co-presence of visual items in the target picture (with *cat* and *fish* going more naturally together than *cat* and *carrots*) should have benefited RW-contexts in eliciting a visual bias towards RW-consistent referents. However, only CW contexts (but not RW contexts) were found to elicit a significant bias towards contextually consistent (CW) referents during picture preview. This supports a model where comprehenders are using the counterfactual discourse to construct a novel 'alternative world', thus relying on higher-level comprehension processes rather than just low-level priming.

Finally, the priming issue has been explicitly addressed in a previous study (Ferguson & Sanford, 2008) to ensure that information in the CW context sentence was not priming readers' access to the critical word in the critical sentence (i.e., *carrots* being primed by *vegetarians*). In this reading study (using sentence materials comparable to the present ones), RW context sentences contained the same potential prime word as CW context sentences, but in a 'realistic' framework, as in '*Evolution dictates that cats are carnivores and cows are vegetarians*'. Using such a modified design, readers showed the same effect patterns as with 'standard' RW contexts (e.g., *If cats are hungry ...*), demonstrating that biases towards CW referents are very unlikely to be purely lexically driven.

Consistent with previous visual-world findings, the relevant 'language-appropriate' referent was visually favoured shortly after the critical word became available in the spoken input. Analyses on how this bias towards language-supported referents developed over time (see Table 3) suggested

that the integration of real-world consistent information (e.g., *Families could feed their cat a bowl of fish* . . .) had a temporal advantage over the integration of real-world inconsistent information (e.g., *Families could feed their cat a bowl of carrots* . . .), regardless of context. This is in line with findings from reading which showed that real-world violations lead to context-independent disruption effects as soon as the critical word is encountered (Ferguson & Sanford, 2008).

In conclusion, the present visual world data point to an important distinction between contextually driven expectation effects on the one hand (which, in this form, cannot be measured in a reading task) and bottom-up word integration effects on the other. We will return to this point in the general discussion.

EXPERIMENT 2

The question arises as to whether context-dependent anticipation effects can be replicated for the similar case of predicting events according to the beliefs of others. It should if people are able to ignore or effectively exclude their knowledge of reality to adopt an alternative 'reality', consistent with the beliefs of others, as the basis of processing. Note that, here, 'reality' refers to a state of affairs portrayed as reality though the narrative, rather than reality inferred from general world knowledge. An example of a *false* belief statement is shown in (2) where reality and the beliefs of another person are in direct conflict with one another.

(2) John washed the dishes after his breakfast and left his watch on the table. While John was distracted, Victoria moved the watch from the table to the bed. Later, John wanted to find his watch so he looked on the bed and yawned.

In this example, context suggests that John is unlikely to know that the watch has moved from the table to the bed (he was *distracted* while that happened), so his reported actions (*he looked on the bed*) are actually inconsistent with his beliefs. In Experiment 2 we investigated such processing of the beliefs of others (theory of mind, ToM). Participants heard a 'reality' or 'belief' context, followed by a target sentence paired with visually presented referents. Eye movements around the visual scene were monitored and time-locked to related auditory input to examine context effects on anticipation towards forthcoming linguistic reality or belief referents. Following from the mental model theory (see introduction), it seems plausible to presume that ToM tasks engage a dual-stage comprehension process, similar to what has been proposed for counterfactuals, where mental spaces are created to represent information about both our own reality and 'reality' according to another person's beliefs. The aim was to allow a fuller investigation into the representation and processing of information within a

specified context. Additionally, we hoped to explore whether the beliefs of others are processed in the same way as counterfactuals, and specifically whether there is a different pattern or time-course of prediction for counterfactual and theory of mind reasoning.

Method

Stimuli and design. Twenty-four sets of three experimental pictures were paired with an auditory passage in one of four conditions. Table 4 and Figure 4 provide an example of such experimental sentences and the associated visual scenes.

The experimental design was similar to that in Experiment 1. Sound files consisted of three sentences: Sentence one introduced a character and described that character putting a target object in a given location. Sentence two then described a second character moving the target object to a new location. This action was either 'explicitly observed' or 'missed' by the first character, creating a 'reality' or a 'belief' context (e.g., '*Later, John noticed Victoria move the* ...', versus '*While John was distracted, Victoria moved the* ...'). A final third sentence drew reference to a *reality*- or *belief*-relevant location (e.g., '*Later, John wanted his watch so he looked on the bed* versus *table* ...'), resulting in a 2×2 within subjects design. Note that the *reality*- or *belief*-language input variables in sentence two refer to consistency with the correspondingly named context. Thus, a *reality*-referent is congruent with a reality context but anomalous in a belief context; equally,

TABLE 4
Examples of experimental sentences (Experiment 2)

Reality context – Belief language input
John washed the dishes after his breakfast and left his watch on the table.
Later, John noticed Victoria move the watch from the table to the bed.
Later, John wanted to find his watch so he looked on the table and yawned.
Reality context – Reality language input
John washed the dishes after his breakfast and left his watch on the table.
Later, John noticed Victoria move the watch from the table to the bed.
Later, John wanted to find his watch so he looked on the bed and yawned.
Belief context – Belief language input
John washed the dishes after his breakfast and left his watch on the table.
While John was distracted, Victoria moved the watch from the table to the bed.
Later, John wanted to find his watch so he looked on the table and yawned.
Belief context – Reality language input
John washed the dishes after his breakfast and left his watch on the table.
While John was distracted, Victoria moved the watch from the table to the bed.
Later, John wanted to find his watch so he looked on the bed and yawned.

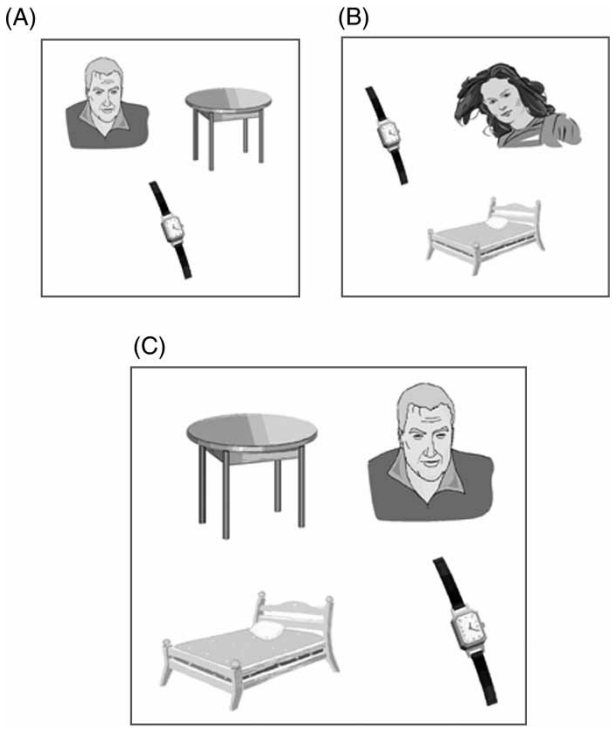

Figure 4. Example visual scenes used in Experiment 2. Participants heard sentence 1 with scene (A), and context sentence 2 with scene (B). Eye-movements were monitored during the target sentence 3 whilst viewing scene (C).

a *belief*-referent is congruent in a belief context and anomalous in a reality context. Three different visual scenes were composed to accompany each auditory sentence. The first scene contained an image of the target object, Character 1 and Location 1. The second scene displayed the target object, Character 2 and Location 2. Finally, the target scene contained four objects: target object, Character 1, reality referent (Location 1), and belief referent (Location 2). The different theory of mind scenarios were set up via auditory context such that the main protagonist (Character 1) either *knew* that the target object has been moved (reality context) or not (belief context). One version of each item was assigned to one of four presentation lists, with each list containing 24 experimental items, 6 in each of the four conditions, blocked to ensure that they were evenly distributed. In addition, 24 filler items were added to each list. They all consisted of correctly matched picture–sentence pairings and were interspersed randomly among the 24 experimental trials to create a single random order. Each subject only saw

each target sentence once, in one of the four conditions. At least one filler trial intervened between any two experimental trials.

Sentences were recorded by the same female native British English speaker as in Experiment 1 and were presented to participants via the same apparatus.

As in Experiment 1, comprehension questions followed half of the experimental and half of the filler trials. The questions could either refer to aspects of the previously presented pictures or to the content of the spoken sentences. Participants did not receive feedback for their responses. All participants scored at or above 90% accuracy on the comprehension questions.

Procedure. The eye tracking procedures were similar to those in Experiment 1. Participants were given the following instruction: 'In this experiment you will hear short spoken passages and during each sentence, a picture will be displayed. We are interested in how the pictures help you understand the spoken passages.'

As illustrated in Figure 5, each trial began with the presentation of a single centrally located dot and participants were asked to fixate it so that automatic drift corrections could be performed. Following successful fixation, the experimenter pressed a button to initiate the trial. The dot was then replaced by Scene 1 while participants heard Sentence 1. Next, Scene 2 was presented with Sentence 2 (reality or belief context). Finally, the target Scene 3 appeared with Sentence 3. A 100 ms blank screen occurred in between any two successive scenes per trial. The onset of each picture preceded the onset of the corresponding spoken sentence by 1000 ms and participants' eye-movements were only recorded during the final (target) picture/sentence presentation. Each trial was automatically ended after 9 seconds; auditory sentences typically ended around 1–2 seconds before the end of the corresponding picture presentation.

As in Experiment 1, the eye-tracker was calibrated and validated at the beginning of each session and once every ten trials thereafter.

Results and discussion

Data processing. Analysis procedures were largely the same as in Experiment 1. Eye-tracking data collected during the final (target) picture/sentence presentation were summarised in terms of the following log-ratio measure:

$$\log(\mathrm{R}/\mathrm{B}) = ln(P_{(\mathrm{R})}/P_{(\mathrm{B})}), \tag{2}$$

where $P_{(\mathrm{R})}$ refers to the probability of gazes on the 'Reality-referent' (the *bed* in our example) and $P_{(\mathrm{B})}$ to the probability of gazes on the 'Belief-referent'

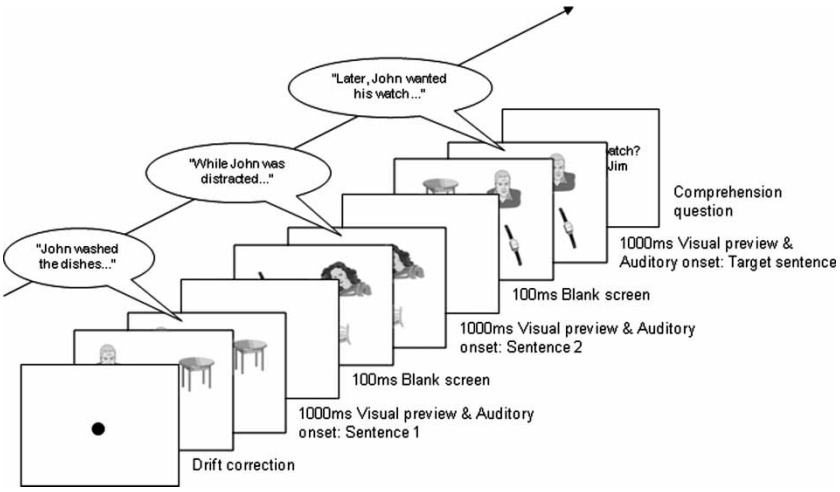

Figure 5. Illustration of the experimental procedure in Experiment 2.

(*table*); positive scores on this measure indicate a visual preference for the Reality-referent and negative scores a visual preference for the Belief-referent.[7]

Since the critical word ('bed' or 'table', respectively) was located closer to the end of the sentence than the critical word in Experiment 1 (two versus seven words prior to sentence conclusion respectively), we analysed a slightly shorter time period in Experiment 2, this time spanning from 1500 ms before critical word onset until 1500 ms after the critical word onset. The critical word period (measured from the onset of the critical word to the onset of the subsequent word) averaged 655 ± 39 ms and 587 ± 38 ms (95% CIs by items) for the 'reality' and 'belief' language input conditions, respectively. Figure 6 shows the corresponding average log(R/B) data per condition, sampled in 20 ms resolution. As in Experiment 1, the solid black vertical line in the figure ($t = 0$) indicates the critical word onset, the dashed lines represent cluster boundaries and the arrow indicates the average verb onset (*looked*). Again, we employed *k*-means cluster analysis to identify larger time windows for analysis. The procedure identified 12 clusters of contiguous 20 ms time slots with similar data patterns across conditions, as shown in the figure. Note that cluster 4 and 5 already happened to border right at the critical word

[7] Again, we found no differential effects of experimental condition on proportions of looks to the remaining referents in the pictures (e.g., *John* or the *watch*), justifying our approach of focusing on Reality- and Belief-referents only. See http://www.langsci.ucl.ac.uk/research/projects/ferguson.pdf for raw probability figures (separately for critical and non-critical objects).

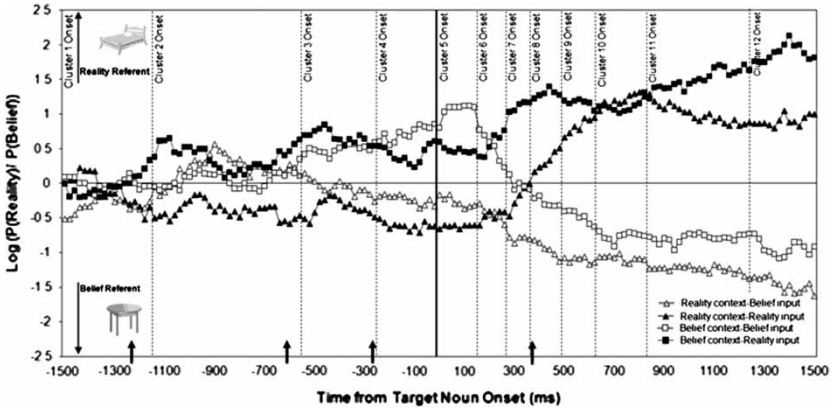

Figure 6. The average log(R/B) as a function of each condition in Experiment 2. Note that the solid black vertical line in the figure (t = 0) indicates the target noun onset, dashed lines represent cluster boundaries for statistical analysis and the arrows indicate (from left to right) the average object offset, verb onset and offset (e.g., looked), and the average target noun offset.

onset. The following inferential analyses were based on weighted average log(R/B) scores per cluster.[8]

Main analyses. For each analysis cluster, we performed a 2×2 ANOVA comprising context (reality versus belief) and language input (reality versus belief) as repeated-measures factors. Table 5 displays the statistical details of the effects, allowing generalisation to participants (F_1) and items (F_2), for each time window of interest.

The analyses revealed no significant effects before cluster 3 (560 to 260 ms prior to critical word onset) where a reliable main effect of context emerged: comparable to Experiment 1, anticipatory fixations were more likely to be made towards contextually relevant referents.[9] Thus, from 560 ms prior to critical word onset, participants were already able to predict events according to the previously induced reality or the beliefs of others. This effect lasted into cluster 4, from 260 ms before to 0 ms (critical word onset), cluster 5 (0–160 ms) and cluster 6 (160–280 ms) with no additional effects of, or interactions with, language input. Also, there was no indication of a general bias in the log-ratio scores (Fs < 1 for the ANOVA *intercepts* not shown in the table for the sake of space). Thus, up until cluster 6, it appears that contextually inconsistent critical words are not being recognised, and that

[8] Figure B in the Appendix shows the alternative word-based clustering of the time series data.

[9] Note that this cluster roughly relates to the verb region using a word-based analysis approach. See Figure B and Table B in the Appendix for statistical results using this approach.

TABLE 5
Analysis of variance results for each time window of interest (Experiment 2)

	F_1				F_2			
Source of variance	*df*	*F_1 value*	*p-value*	*$_p\eta^2$*	*df*	*F_2 value*	*p-value*	*$_p\eta^2$*
(−1500)–(−1140) ms								
Context	1, 27	1.7	.2	0.06	1, 23	0.02	.9	0.9
Language input	1, 27	0.005	.94	0	1, 23	0.33	.57	0.57
Context × Language input	1, 27	0.13	.72	0.01	1, 23	1.02	.32	0.32
(1140)–(−560) ms								
Context	1, 27	I.82	.19	0.06	1, 23	2.04	.17	0.17
Language input	1, 27	1.26	.27	0.05	1, 23	0.14	.71	0.7
Context × Language input	1, 27	3.8	.06	0.12	1, 23	4.23	.16	0.05
(−560)–(−260) ms								
Context	1, 27	11.67	.002***	0.3	1, 23	17.36	<.001"*	0.43
Language input	1, 27	0.36	.56	0.01	1, 23	0.35	.57	0.02
Context × Language input	1, 27	0.d23	.63	0.01	1, 23	1.82	.19	0.07
(−260)–0 ms								
Context	1, 27	10.53	.003***	0.28	1, 23	19.49	<.001***	0.46
Language input	1, 27	0.52	.4S	0.02	1, 23	2.93	.1	0.11
Context × Language input	1, 27	0.002	.96	0	1, 23	2.07	.16	0.08
0–160 ms								
Context	1, 27	15.87	<.001***	0.37	1, 23	26.95	<.001***	0.54
Language input	1, 27	1.7	.2	0.06	1, 23	3.8	.06	0.14
Context × Language input	1, 27	0.24	.63	0.01	1, 23	0.04	.84	0
160–280 ms								
Context	1, 27	14.24	.001***	0.35	1, 23	16.3.9	<.001***	0.42
Language input	1, 27	0.14	.71	0.01	1, 23	0.07	.79	0
Context × Language input	1, 27	0.03	.86	0	1, 23	0.01	.92	0

Continued

TABLE 5 (*Continued*)

	F_1				F_2			
Source of variance	*df*	F_1 *value*	*p-value*	$_p\eta^2$	*df*	F_2 *value*	*p-value*	$_p\eta^2$
280–380 ms								
Context	1, 27	18.11	<.001***	0.4	1, 23	22.51	<.001***	0.5
Language input	1, 27	13.77	.001***	0.34	1, 23	16.42	<.001***	0.42
Context × Language input	1, 27	1.03	.32	0.04	1, 23	0.9	.35	0.04
380–500 ms								
Context	1, 27	9.49	.005***	0.26	1, 23	17.66	<.001"*	0.43
Language input	1, 27	38.34	<.00l***	0.59	1, 23	43.92	<.001***	0.66
Context × Language input	1, 27	0.37	.55	0.01	1, 23	0.01	.91	0
500–640 ms								
Context	1, 27	2.61	.12	0.09	1, 23	6.16	.02*	0.21
Language input	1, 27	54.45	<.001***	0.67	1, 23	51.07	<.001"*	0.69
Context × Language input	1, 27	0.2	.66	0.01	1, 23	0.39	.35	0.04
640–860 ms								
Context	1, 27	1.18	.29	0.04	1, 23	1.11	.3	0.05
Language input	1, 27	72.31	<.001***	0.72	1, 23	73.63	<.001***	0.76
Context × Language input	1, 27	1.21	.28	0.04	1, 23	1.95	.18	0.08
860–1260 ms								
Context	1, 27	7.04	.01**	0.21	1, 23	5.97	.02*	0.21
Language input	1, 27	40.28	<.001***	0.61	1, 23	52.06	<.001***	0.69
Context × Language input	1, 27	0.04	.34	0	1, 23	0.23	.63	0.01
1260–1500 ms								
Context	1, 27	8.38	.007**	0.24	1, 23	8.12	.003**	0.27
Language input	1, 27	41.76	<.001***	0.61	1, 23	49.99	<.001***	0.69
Context × Language input	1, 27	2.13	.16	0.07	1, 23	0.74	.4	0.03

Note: *$p < .05$; **$p < .01$; ***$p < .001$.

context is the sole determinant of participants' anticipatory eye-movement behaviour.

However, from cluster 7 (280–380 ms) a main effect of language input emerged, revealing that from 280 ms after critical word onset, participants visually favoured the relevant 'appropriate' referent. This effect appeared alongside a reliable context main effect. Cluster 8 (380–500 ms) showed a similar pattern of results, with reliable effects of language input and context. Clusters 9 (500–640 ms) and 10 (640–860 ms) maintain the significant effect of language input. However, effects of context have largely faded away in these time windows. Clusters 11 (860–1260 ms) and 12 (1260–1500 ms) also showed a reliable main effect of language input, and interestingly, reliable effects of context re-emerged in these time windows. Thus, it appears that prior context regarding reality or the beliefs of others has a very early (anticipatory) and enduring influence on language comprehension in this kind of paradigm.

Finally, note that there was no clear suggestion of language-input dependent differences in how visual biases towards 'appropriate' referents developed over time, contrasting with Experiment 1. To confirm this, we conducted statistical analyses of cluster ranks with language input (Reality vs. Belief, averaged across the two context conditions) and cluster (7, 8, 9, 10, 11, 12) as repeated-measures factors. Mean cluster ranks (scored from 1 to 6) are shown in Table 6, together with 95% confidence limits both for the rank means and for the Reality-Belief rank differences per cluster. In sum, the two-way ANOVAs revealed a main effect of cluster, $F_1(5, 135) = 6.08$, $p = .001$; ${}_p\eta 2 = 0.57$; $F_2(5, 115) = 5.49$, $p < .005$; ${}_p\eta 2 = 0.59$, reflecting the differing average ranks across the six clusters. More importantly, there was no interaction between cluster and language input (Fs < 1.27), thus demonstrating that integration of 'input-appropriate' referents is not influenced by temporarily established reality or beliefs information.

Preview region. As in Experiment 1, anticipation effects were examined in the preview period prior to linguistic input. Specifically, we analysed the time window from 500 ms to 1000 ms post-picture onset, as before. Once again, this time window did not overlap with the main analysis time window for any experimental items. Statistical analyses were carried out on the probabilities of gazes to the critical Reality and Belief referents as a function of time, using the same log-ratio measure as explained previously (Eq. 2). Inferential analyses revealed no significant main effect of context or language input, and there was also no reliable context by language input interaction (all Fs < 2.8).

In sum, Experiment 2 showed that participants were, on average, able to use information from a prior context to update their knowledge about reality and the beliefs of others to anticipate a relevant referent from 560 ms prior to

TABLE 6
Mean cluster ranks representing ordinal strength of bias towards 'appropriate' referents per cluster as a function of language input (Reality vs. Belief congruent). Ninety-five per cent confidence limits (by participants and items) are provided for the cluster ranks themselves as well as for the Reality-Belief input difference per cluster.

	Cluster rank (by subjects)	*Cluster rank (by items)*
Cluster 7 (280–380 ms)		
Reality input	2.68 ± 0.33	2.25 ± 0.31
Belief input	2.82 ± 0.30	3.00 ± 0.37
Difference	−0.14 ± 0.24	−0.75 ± 0.24
Cluster 8 (380–500 ms)		
Reality input	2.86 ± 0.26	2.79 ± 0.29
Belief input	2.46 ± 0.24	2.71 ± 0.34
Difference	+0.40 ± 0.21	+0.08 ± 0.21
Cluster 9 (500–640 ms)		
Reality input	3.54 ± 0.34	3.58 ± 0.31
Belief input	3.14 ± 0.28	3.25 ± 0.27
Difference	+0.40 ± 0.24	+0.33 ± 0.21
Cluster 10 (640–860 ms)		
Reality input	3.89 ± 0.31	3.71 ± 0.36
Belief input	4.43 ± 0.27	4.04 ± 0.28
Difference	−0.54 ± 0.19	−0.33 ± 0.21
Cluster 11 (800–1260 ms)		
Reality input	4.07 ± 0.32	3.83 ± 0.37
Belief input	4.07 ± 0.33	4.17 ± 0.34
Difference	0.00 ± 0.26	−0.34 ± 0.28
Cluster 12 (1260–1500 ms)		
Reality input	4.00 ± 0.30	4.17 ± 0.38
Belief input	4.11 ± 0.35	4.50 ± 0.28
Difference	−0.11 ± 0.24	−0.33 ± 0.23

the critical word onset. This supports findings from Experiment 1 that prior context is rapidly processed so that participants *expect* a context-relevant continuation. As with counterfactuals, the relevant 'language-appropriate' referent was visually favoured shortly after the critical (reality or belief-referent) word became available in the sentence. However, contrasting with Experiment 1, there was no difference in the dynamics of integrating reality or belief-consistent language input. This makes sense because the establishment of a reality or belief-context in Experiment 2 did not draw upon *pre-stored* world knowledge information, unlike the stereotypical/counterfactual relationships investigated in Experiment 1.

Another contrast to Experiment 1 was that no effects of context were observed in the preview region. This suggests that while people can mentally

represent both reality and an alternative 'reality' based on another person's beliefs, they do not build up expectations of forthcoming events according to this knowledge until necessary. This suggests that despite the structural similarity between counterfactual and false belief statements, different cognitive processes are involved in their comprehension. These issues will be discussed in full below.

GENERAL DISCUSSION

Previous experiments have shown that discourse processing is typically driven by predictive relationships involving syntax, semantics and real-world expectations. Therefore, the primary issue investigated in the present paper was whether and to what extent the constraints of real-world knowledge and prior context influence eye-movements directed towards objects in the visual world. Counterfactuals have recently been investigated using eye-tracking in reading (Ferguson & Sanford, 2008) where results suggested delayed integration effects from the counterfactual context following initial interference from real-world knowledge. In comparison, the issue of on-line theory of mind processing thus far has been largely neglected. As such, very little is known about the on-line processes that are activated when comprehenders draw inferences on a ToM scenario. In this paper, we have explored whether people can use their knowledge of the wider discourse (specifically using either a counterfactual or beliefs of others framework) to override real-world knowledge to predict specific upcoming words as the current sentence is unfolding.

Anticipation based on context

Experiment 1 used the visual world paradigm to investigate whether the typical anticipatory bias towards real-world consistent objects in this task (cf. *little girls are more likely to ride carrousels than motorbikes* in Kamide et al., 2003 versus *cats are more likely to eat fish than carrots* in the present study) could be eliminated by introducing an appropriate counterfactual context of the form, 'If X were Y then . . .'. The results showed that anticipatory eye-movements were made towards contextually relevant referents from at least 200 ms prior to the critical word onset. The fact that participants direct their attention towards real-world relevant objects in a concurrent visual scene following a real-world context is not surprising given previous research suggesting such anticipation. However, the fact that this real-world bias can be temporarily suppressed in favour of a counterfactual-world relevant referent is a novel observation with theoretical implications, which will be discussed shortly.

Experiment 1 also revealed evidence of very early prediction based on the discourse context, with increased looks towards the CW referent during picture preview following a CW context. However, following a RW context, no such early anticipation towards the RW referent was evident. This suggests that during the CW context sentence, participants are already forming assumptions as to an appropriate continuation, and when presented with a limited set of visual referents, they are able to make rapid predictions about appropriate continuations. But why is this not the case following a RW context? We suggest that participants were more susceptible to creating expectations according to the CW context because reality in the RW context was 'implied', whereas the CW context explicitly stated a hypothetical counterfactual scenario. This would account for the significant anticipation towards the contextually relevant referent even before target sentence onset. In contrast, following the RW context, participants delay their expectations to seek more information from the upcoming linguistic input. It is also interesting to note that participants do not appear to anticipate negative continuations in these scenarios (e.g., 'If cats were vegetarians, they wouldn't eat fish'). However, as has been demonstrated by several studies of negation (e.g., Ferguson, Sanford, & Leuthold, 2008; Kaup, Lüdtke, & Zwaan, 2006), representing a negative situation is frequently more difficult to process and is therefore subject to a processing delay (but see Nieuwland & Kuperberg, 2008), which may explain why affirmative antecedent continuations are the preferred choice with the stimuli used here.

Further evidence for anticipation according to prior contexts was found in Experiment 2, where a prior context directed comprehenders to make sense of the passage either according to narrative reality or, more interestingly, according to the beliefs of another person. Effects similar to Experiment 1 were found, with typical real-world biases, based on information about the Agent and the verb, temporarily eliminated by an appropriate 'beliefs of others' prior context. Anticipatory eye-movements were made towards contextually relevant referents from 560 ms prior to the critical word onset. In other words, context was able to elicit very early expectancy effects that modified the constraints of narrative reality. So the present studies provide evidence for an incremental language processor that makes immediate use of all information available to construct a plausible interpretation of the linguistic input, as suggested by Kamide et al. (2003). However, the current studies expand on this suggestion as we demonstrate that linguistic context can overrule experiential or narrative-based knowledge of objects and their interactions. Thus, comprehenders are able to create *novel* relationships between objects in the scene as the sentence unfolds. In short, monitoring eye-movements around a visual scene can reveal expectations established either through real-world knowledge (e.g. Altmann & Kamide, 1999; Kamide

et al., 2003) or from some appropriate alternative linguistic input such as counterfactual context (Experiment 1) or theory of mind (Experiment 2).

The results from Experiments 1 and 2 are consistent with the development of 'mental spaces' to represent information during language comprehension. Within the mental model theory of conditionals, Santamaria, Espino, & Byrne (2005) suggested that a counterfactual conditional statement creates both 'factual' and 'counterfactual' possibilities whereas a factual conditional creates only the 'real-world' possibility. Similarly, we can apply the mental model theory to the comprehension of the 'beliefs of others', which creates both 'reality' and 'belief' spaces. The fact that participants in these experiments were quickly able to use contextual information, even in cases where doing so conflicted with real-world knowledge or *reality*, could be taken to support the view that comprehenders are only representing this one possibility. However, as demonstrated by the counterfactual statements in Experiment 1 (see also Ferguson & Sanford, 2008), integration of language input is affected by world knowledge at the point of a violation. As such, this provides evidence that although comprehenders readily accept this counterfactual scenario, integration of events within that context remains grounded in factual knowledge. Interestingly, such bottom-up integration effects were not evident for comprehending the beliefs of others in Experiment 2. We propose that the lack of integration effects in this ToM experiment is due to the fact that the counterfactuals under investigation here involved pre-stored world knowledge (note that this assertion does not apply to counterfactuals generally; consider, e.g., 'If it were sunny outside, I would wear sunglasses'). In contrast, the ToM stories manipulated *temporarily* established 'reality' or 'belief' situations, which do not elicit such strong anomaly detection responses (general world knowledge is indifferent as to whether '*John notices that Victoria moved the watch from the table to the bed*', for example). Taken together, this experimental evidence strongly supports the theoretical suggestion that two possibilities are represented during counterfactual and ToM comprehension (although the impact of these representations is substantially weaker for the temporarily established beliefs of others).

Context grounded in reality?

It is interesting to note that no interference from real-world knowledge was evident in the *anticipatory* eye-movement results from Experiments 1 or 2. This is a novel finding suggesting that top-down expectations during processing appear to be solely driven by context. Eye-tracking studies on reading, by contrast, primarily tapped into the cost of *integrating* RW-consistent or -inconsistent information. For example, Ferguson and Sanford (2008) demonstrated that RW-inconsistent language input leads to early processing disruption, regardless of context. Indeed, Experiment 1 in the

present paper revealed findings that corroborate this conclusion: upon recognising the critical word in the auditory language input, participants' visual biases towards 'appropriate' referents developed faster with RW-consistent rather than RW-inconsistent language input. However, it is important to bear in mind that these effects emerged *after* the critical word has been available to the listener, just as findings from reading were established *after* the critical word has been fixated for the first time. Hence, we propose that it is important to distinguish between bottom-up integration of linguistic input on the one hand and top-down prediction of forthcoming input on the other. While pre-stored world knowledge and narrative reality appear to have an influence on the former, the latter is predominantly context-driven. Importantly, knowledge of the real world still bears upon this contextualised interpretation (i.e., that vegetarians would eat carrots), as predicted by Frith's (1989) theory of Strong Central Coherence (also see Fauconnier, 1985, 1997). Thus, top-down predictions can be established through combinations of verb-based information (e.g., *eat*), a pre-verbal argument (Agent) (e.g., *the cat*) and as demonstrated here, an appropriate discourse context. The visual world paradigm encourages the formation of predictions since the visual referents prompt the comprehender to incrementally assess the fit of the referents to the current linguistic input. In contrast, no such predictive cues are available in a reading task, meaning that anticipation, though surely occurring to some extent, cannot be measured in the corresponding eye-movement records. In conclusion, we suggest that the different tasks complement each other and that combining results across paradigms leads to a fuller understanding of language processing than when used individually.

In line with the data reported here, two processes became manifest. The first process in the comprehension of counterfactuals is to create expectations about the unfolding discourse according to a contextually updated model of the world. This stage is particularly evident given the very early context-based anticipation in the preview window in Experiment 1. Second, newly encountered input is briefly checked against pre-stored world knowledge. This stage becomes apparent in effects occurring *after* the critical word in the language input has become available to the reader or listener: both reading and visual-world data demonstrate an early, time-limited conflict arising upon encountering RW-inconsistent input, regardless of context. Hence, it appears that context-consistent (i.e., anticipation) and real-world consistent (i.e., integration) mental models are represented in parallel during the processing of counterfactuals.

Clearly an interesting issue for further research is to investigate the structure of mental representations that embody counterfactual and ToM situations (including timing and relative availability of individual elements). Traditionally, the literature has pointed to an egocentric advantage in ToM

situations, where adults design and interpret utterances from their own perspective, only adjusting to others' perspectives when they make an error (e.g., Keysar, Barr, Balin, & Brauner, 2000; Keysar, Lin, & Barr, 2003; see Barr & Keysar, 2007 for a detailed review). However, the results reported here contradict this suggestion and instead support a model where ToM (and similarly, counterfactual) processing has an early effect on language comprehension without initial interference from narrative reality. This evidence is in line with data from Hanna et al. (2003); see also Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Hanna & Tannenhaus, 2004; Metzing & Brennan, 2003; Nadig & Sedivy, 2002) who have collectively used a variety of sentence structures to demonstrate communicators' ability to make a clear distinction between shared and private information, often resulting in very early perspective-based referential biases during language processing. Thus, an important process in the comprehension of ToM and counterfactuals is to create expectations about unfolding events according to a contextually updated model of the world. This finding is synonymous with previous investigations of discourse context, which show that context can have a very early influence on the interpretation of different linguistic constructs (e.g., Altmann & Kamide, 1999; Ferguson & Sanford, 2008; Filik, 2008; Hess, Foss, & Carroll, 1995; Nieuwland & Van Berkum, 2006; Pickering & Traxler, 1998).

Counterfactuals and theory of mind as related processes

In reference to the proposed link between understanding counterfactuals and beliefs of others (Perner, Sprung, & Steinkogler, 2004; Peterson & Bowler, 2000; Riggs et al., 1998), the experiments reported here demonstrate mixed evidence for the recruitment of related processes. Rapid contextually driven anticipation effects in both studies support the involvement of similar mental models to achieve a full understanding of the linguistic input. This is likely to reflect that fact that both tasks require comprehenders to create two mental spaces to represent the linguistic information provided. In the case of counterfactual conditionals, one is the factual and the other is the counterfactual hypothetical space, whereas theory-of-mind tasks require reality and the beliefs of others to be represented in separate spaces. However, the relative availability of information in these two mental spaces appears to differ between counterfactual and beliefs of others tasks.

This claim is supported by the fact that within a counterfactual context, participants initiated contextually relevant predictions about the subsequent continuation during the one-second visual 'preview' period, prior to auditory sentence onset. By contrast, beliefs of others contexts did not elicit such early anticipatory eye-movements during preview. This suggests that within a counterfactual discourse of the form 'If X were Y then . . .' participants

append some appropriate consequence of the counterfactual world to their mental model of that context. In contrast, although people can mentally represent both reality and an alternative 'reality' based on another person's beliefs, generating expectations of forthcoming events (including predicting whose perspective to take) is clearly a more abstract task than the concrete events described by counterfactuals. Accordingly, comprehenders delay their expectations until prompted by later linguistic input. Further investigations are essential to gain a detailed understanding of the specific and overlapping encoding and processing mechanisms employed in these tasks.

In conclusion, visual attention can be immediately directed according to the expectations constructed from a prior discourse context, thus influencing how we anticipate the consequences of the unfolding world. The studies reported here provide novel evidence that this is true even if those discourse constraints are inconsistent with real-world expectations that normally determine which referents are appropriate given local semantic restrictions. Furthermore, the current results taken with previous eye-tracking reading results (Ferguson & Sanford, 2008) emphasise the benefits of employing both reading and visual-world paradigms to obtain a full understanding of the processes underlying language comprehension, incorporating both integration costs and developing predictions.

Manuscript received August 2008
Revised manuscript received May 2009
First published online Month/year

REFERENCES

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Altmann, G.T.M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*, 55–71.

Arai, M., van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*, 218–250.

Arnold, J. E., Wasow, T., Asudeh, A., & Alrenga, P. (2004). Avoiding attachment ambiguities: The role of constituent ordering. *Journal of Memory and Language*, *51*, 55–70.

Baron-Cohen, S. (2000). Autism: deficits in folk psychology exist alongside superiority in folk physics. In S. Baron-Cohen, H. Tager Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism and developmental cognitive neuroscience* (2nd ed.). Oxford, UK: Oxford University Press.

Barr, D. J., & Keysar, B. (2007). Perspective taking and the coordination of meaning in language use. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 901–938). London: Academic Press.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*, 1122–1134.

Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Science*, *6*, 426–431.

Byrne, R. M. J. (1997). Cognitive processes in counterfactual thinking about what might have been. In D. L. Medin (Series Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 37, pp. 105–154). San Diego, CA: Academic Press.

Byrne, R. M. J., & Tasso, A. (1999). Deductive reasoning with factual, possible, and counterfactual conditionals. *Memory and Cognition*, *27*, 726–740.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Epley, N., Morewedge, C., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology*, *40*, 760–768.

Fauconnier, G. (1985). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge, MA: MIT Press.

Fauconnier, G. (1997). *Mappings in thought and language*. New York: Cambridge University Press.

Fauconnier, G., & Turner, M. (2003). *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.

Ferguson, H. J., & Sanford, A. J. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, *58*, 609–626.

Ferguson, H. J., Sanford, A. J., & Leuthold, H. (2008). Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research*, *1236*, 113–125.

Ferguson, H. J., Sanford, A. J., & Leuthold, H. (2006). Detecting violations in real- and counterfactual-world contexts: Eye-movements and ERP analysis. Paper presented at *Architectures and Mechanisms for Language Processing*. Nijmegen, the Netherlands.

Ferguson, H. J., Sanford, A. J., & Scheepers, C. (2008). On-line investigations of theory of mind reasoning. Poster presented at the *Workshop on Pragmatics and Social Cognition*. University College London, UK.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*, 296–340.

Filik, R. (2008). Contextual override of pragmatic anomalies: Evidence from eye movements. *Cognition*, *106*, 1038–1046.

Frith, U. (1989). *Autism: Explaining the enigma*. Oxford, UK: Blackwell.

Frith, C. D., & Corcoran, R. (1996). Exploring 'theory of mind' in people with schizophrenia. *Psychological Medicine*, *26*, 521–530.

Frith, C. D., & Frith, U. (1988). Electives affinities in schizophrenia and childhood autism. In P. Bebbington (Ed.), *Social psychiatry: Theory, methodology and practice* (pp. 65–88). New Brunswick, NJ: Transactions Publishers.

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Science*, *7*, 77–83.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science*, *28*, 105–115.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*, 43–61.

Happé, F., Malhi, G. S., & Checkley, S. (2001). Acquired mindblindness following frontal lobe surgery? A single case study of impaired 'theory of mind' in a patient treated with stereotactic anterior capsulotomy. *Neuropsychologia*, *39*, 83–90.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS136: A k-means clustering algorithm. *Applied Statistics*, *28*, 100–108.

Haywood, S. L., Pickering, M. J., & Branigan, H. P. (2005). Do speakers avoid ambiguities during dialogue? *Psychological Science*, *16*, 362–366.

Hess, D. J., Foss, D. J., & Carroll, P. (1995). Effects of global and local context on lexical decision processing during language comprehension. *Journal of Experimental Psychology: General*, *124*, 62–82.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, *96*(1), 23–32.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, UK: Cambridge University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Kahneman, D., & Miller, D. T. (1986). Norm theory – comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133–159.

Kamide, Y., Scheepers, C., & Altmann, G. T. M. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of Psycholinguistic Research*, *32*(1), 37–55.

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*, 1033–1050.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*, 32–38.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, *89*, 25–41.

Leslie, A. (1987). Pretence and representation: the origins of a 'theory of mind'. *Psychological Review*, *94*, 412–426.

Leslie, A. M. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition*, *50*, 193–200.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*, 1–24.

Markman, K. D., & Tetlock, P. E. (2000). 'I couldn't have known': Accountability, foreseeability and counterfactual denials of responsibility. *British Journal of Social Psychology*, *39*(3), 313–325.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective taking constraints in children's on-line reference resolution. *Psychological Science*, *13*, 329–336.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth isn't too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*, 1213–1218.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, *18*, 1098–1111.

Perner, J., Sprung, M., & Steinkogler, B. (2004). Counterfactual conditionals and false belief: A developmental dissociation. *Cognitive Development*, *19*, 179–201.

Peterson, D. M., & Bowler, D. M. (2000). Counterfactual reasoning and false belief understanding in children with autism. *Autism: The International Journal of Research and Practice*, *4*, 391–405.

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 940–961.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice-Hall.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, *13*, 73–91.

Rowe, A. D., Bullock, P. R., Polkey, C. E., & Morris, R. G. (2001). 'Theory of mind' impairments and their relationship to executive functioning following frontal lobe excisions. *Brain*, *124*, 600–616.

Santamaria, C., Espino, O., & Byrne, R. M. J. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 1149–1154.

Scheepers, C., & Crocker, M. (2004). Constituent order priming from reading to listening: A visual-world study. In M. Carreiras & C. Clifton, Jr. (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERP and beyond*. New York: Psychology Press.

Scheepers, C., Keller, F., & Lapata, M. (2008). Evidence for serial coercion: A time course analysis using the visual-world paradigm. *Cognitive Psychology*, *56*, 1–29.

Sedivy, J., Tanenhaus, M., Chambers, C., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*, 109–147.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, *10*, 640–656.

Stuss, D. T., Gallup, G. G. Jr., & Alexander, M. P. (2001). The frontal lobes are necessary for theory of mind. *Brain*, *124*, 279–286.

Tager-Flusberg, H., Boshart, J., & Baron-Cohen, S. (1998). Reading the windows of the soul: Evidence of domain specificity sparing in Williams syndrome. *Journal of Cognitive Neuroscience*, *10*, 631–639.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Van Berkum, J. J. A. (2008). Understanding sentence in context: What brain waves can tell us. *Current Directions in Psychological Science*, *17*(6), 376–380.

Van Berkum, J. J. A., Holleman, B. C., Murre, J. M. J., Nieuwland, M. S., & Otten, M. (2007). *So how do you feel about this? An ERP study on opinion poll comprehension.* Annual Meeting of the Cognitive Neuroscience Society, New York.

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.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1–14.

Appendix

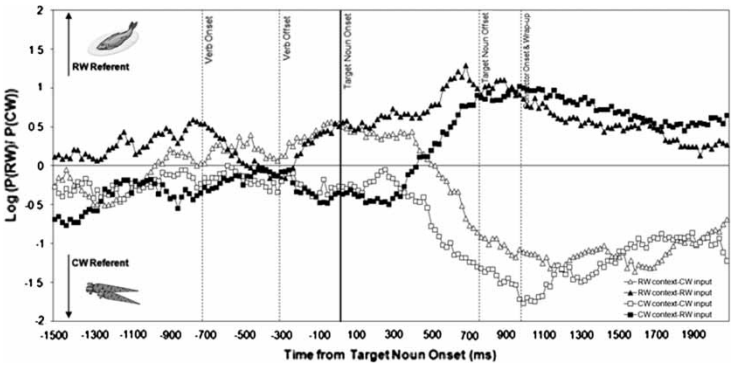

Figure A. Average log(RW/CW) scores as a function of time and condition in Experiment 1. Vertical lines indicate (averaged over trials for illustration purposes) the verb onset and offset, target noun onset and offset, and connector onset and wrap-up.

Determining analysis windows via *k*-means clustering

A common problem in time series data analysis is how to specify time windows of interest. Two approaches are frequently used in the literature. One is to divide the whole time series into k equally sized analysis windows and to perform inference statistical analyses for each of these windows after averaging the data across the constituent time slots per window. The disadvantage of this method is that it often masks potentially important detail. For example, imagine that within a given analysis window and condition, a visual bias towards object A changes into an equally sized bias towards object B; aggregating data over the relevant time slots will not be able to detect this change in visual bias over time. An alternative method for analysing time series data would be to divide the time slots into k unequally-sized analysis windows, based on visual inspection of the data. Again, this is equivalent to a reduction in temporal resolution. However, the researcher can at least make sure that he/she will not miss anything important. (ERP data often used to be analysed in this way.) The problem with this approach is that it can be rather arbitrary and subjective. Moreover, it is not necessarily very precise, specifically if more than two conditions are involved while it appears relatively easy to visually divide data from two conditions into time periods of interest, the task becomes far more challenging with three, four, or even more conditions, especially if these conditions display all sorts of complex interactions over time.

The approach taken in this paper was to use *k-means cluster* analysis (Hartigan, 1975; Hartigan & Wong, 1979) as an auxiliary procedure in identifying time windows of interest. *K-means cluster* analysis employs an iterative sorting algorithm whereby n observations (in this case time slots) are classified into k clusters of observations ($1 < k < n$) such that between-cluster similarity is minimised and within-cluster similarity is maximised, given some classification criteria (a set of variables known as *clustering dimensions*). The clustering dimensions are combined into a single similarity metric (typically based on Euclidian Distance – the shortest distance between two arbitrary points in a multi-dimensional space) by which any two observations can be compared and sorted relative to one another. The number of clusters, k,

TABLE A
Analysis of variance results for each word-based time window of interest (Experiment 1).
The time-windows were defined on an item-by-item basis

	F_1				F_2			
Source of variance	*df*	F_1 *value*	*p-value*	$_p\eta^2$	*df*	F_2 *value*	*p-value*	$_p\eta^2$
Verb								
Context	1, 27	5.11	.03*	0.16	1, 23	0.67	.42	0.03
Language input	1, 27	0.01	.92	0	1, 23	0.12	.74	0.01
Context × Language input	1, 27	0.12	.74	0.01	1, 23	0.73	.4	0.03
Post-verb Break								
Context	1, 27	12.01	.002***	0.31	1, 23	4.84	.03*	0.15
Language input	1, 27	0.39	.54	0.01	1, 23	0.24	.63	0.01
Context × Language input	1, 27	0.37	.55	0.01	1, 23	0.03	.96	0
Target Noun								
Context	1, 27	10.86	.003***	0.29	1, 23	16.92	<.001***	0.42
Language input	1, 27	3.59	.07	0.12	1, 23	7.29	.01**	0.24
Context × Language input	1, 27	0.26	.61	0.01	1, 23	0.05	.83	0
Post-target Noun Break								
Context	1, 27	0.93	.34	0.03	1, 23	1.68	.21	0.07
Language input	1, 27	54.11	<.001***	0.67	1, 23	75.3	<.001***	0.77
Context × Language input	1, 27	1.09	.31	0.04	1, 23	0.01	.91	0
Connector & Wrap-up								
Context	1, 27	0.58	.45	0.02	1, 23	0.08	.79	0
Language input	1, 27	34.33	<.001***	0.56	1, 23	47.08	<.001***	0.67
Context × Language input	1, 27	2.67	.11	0.09	1, 23	0.54	.47	0.02

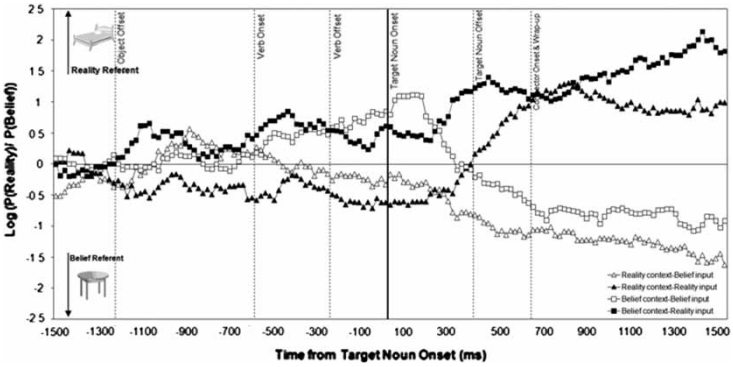

Figure B. Average log(R/B) scores as a function of time and condition in Experiment 2. Vertical lines indicate (averaged over trials for illustration purposes) the object offset, verb onset and offset, target noun onset and offset, and connector onset and wrap-up.

needs to be specified a priori by the user. However, further below we will describe a heuristic whereby the optimum k can be determined.

In our analyses (for technical details, please refer to the second author), the to-be-classified cases consisted of the n time slots considered in each experiment ($n = 175$ in Experiment 1; $n = 150$ in Experiment 2). The clustering dimensions were the four log-ratio grand averages per time slot (one for each experimental condition) and the time variable itself. The inclusion of the latter ensured that adjacent time slots were more likely to end up within the same cluster. Since the similarity metric is scale-sensitive, the time variable and the log-ratio scores were z-transformed before entering the analyses. (Hence, each clustering dimension was given the same weight.)

By assuming $k = 5$, for example, the *k-means clustering* algorithm will identify five time-slot clusters that are (descriptively) maximally distinct in terms of cross-condition data patterns over time, as shown in Figure C. By increasing k, more detail can be resolved.

Before moving on to explain how the optimum number of clusters can be determined, it is important to stress that the procedure is not geared towards maximising cross-condition effect sizes, which could result in biased sampling and thus alpha inflation. To illustrate this point, we performed analyses on an idealised data set with two conditions (A and B) and 24 equally spaced time slots (Figure D). The data in condition A were generated from an asymptotically increasing asymmetric sigmoid function, and the data in condition B from the inverse of that function. As before, the time variable (x-axis) and the data points (y-axis) were z-transformed for analysis. The top panel in Figure D shows the results from a *k-means cluster* analysis over the 24 time slots; k was set to 4, and the relevant condition scores, as well as the time variable, were used as clustering dimensions. The middle panel in Figure D represents an evenly spaced four-cluster partitioning of the whole time series. Finally, the bottom panel in Figure D shows a four-cluster solution whereby the average effect size per cluster is maximized. Also shown in each panel is the average Condition A – Condition B difference (d) per cluster. As can be seen, the *k-means clustering* solution (top panel) differs markedly from the effect-maximisation solution (bottom panel). The average absolute effect size per cluster amounts to $(2.50+0.46+1.39+1.77)/4 = 1.53$ for the *k-means clustering* solution (top panel), to $(2.65+1.80+1.20+1.77)/4 = 1.86$ for the equal spacing solution (middle panel), and to $(2.65+2.36+1.30+1.77)/4 = 2.02$ for the effect-maximisation solution (bottom panel). The latter even exceeds the average absolute effect size obtained over the original 24 time slots (which, for the given data set, is equivalent to that from the equal spacing solution, i.e. 1.86). These results clearly demonstrate that the *k-means*

TABLE B
Analysis of variance results for each word-based time window of interest (Experiment 2).
The time-windows were defined on an item-by-item basis.

	F_1				F_2			
Source of variance	*df*	*F_1 value*	*p-value*	*$_p\eta^2$*	*df*	*F_2 Value*	*p-value*	*$_p\eta^2$*
Post-object Break								
Context	1, 27	1.42	.24	0.05	1, 23	2	.17	0.08
Language input	1, 27	1.4	.25	0.05	1, 23	0.03	.86	0
Context × Language input	1, 27	2.85	.01	0.1	1, 23	2.35	.14	0.09
Verb								
Context	1, 27	11.89	.002***	0.31	1, 23	19.5	<.001***	0.46
Language input	1, 27	0.45	.51	0.02	1, 23	0.39	.54	0.02
Context × Language input	1, 27	0.34	.56	0.01	1, 23	2.01	.17	0.08
Post-verb Break								
Context	1, 27	10.07	.004***	0.27	1, 23	19.25	<.001***	0.46
Language input	1, 27	0.52	.48	0.02	1, 23	3.03	.1	0.12
Context × Language input	1, 27	0	.96	0	1, 23	1.96	.18	0.08
Target Noun								
Context	1, 27	15.23	.001***	0.36	1, 23	22.45	<.001***	0.49
Language input	1, 27	2.24	.15	0.08	1, 23	1.45	.24	0.06
Context × Language input	1, 27	0.23	.64	0.01	1, 23	0.01	.91	0
Post-target Noun Break								
Context	1, 27	4.54	.04*	0.14	1, 23	13.24	.001***	0.37
Language input	1, 27	55.26	<.001***	0.67	1, 23	55.3	<.001***	0.71
Context × Language input	1, 27	0	.97	0	1, 23	0.45	.51	0.02
Connector & Wrap-up								
Context	1, 27	5.5	.03*	0.17	1, 23	4.26	.05*	0.16
Language input	1, 27	50.98	<.001***	0.65	1, 23	54.15	<.001***	0.7
Context × Language input	1, 27	0.28	.6	0.01	1, 23	0.77	.39	0.03

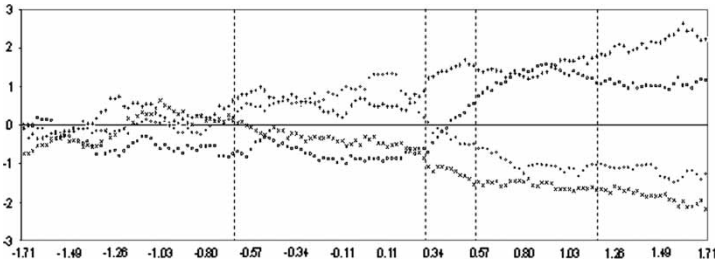

Figure C. Example time series data (taken from Experiment 2) divided into five clusters of time slots (indicated by dashed vertical lines) such that between-cluster similarity is minimised and the within-cluster similarity is maximised. Different marker symbols represent different experimental conditions. Note that the clustering dimensions (the time variable and the by-condition log-ratio scores) were z-transformed for analysis.

clustering procedure is *not* biased towards maximising cross-condition effect sizes per cluster – it is primarily interested in similarity of data patterns over time.[10]

The remaining problem, then, is to determine the minimum number of clusters required to describe the time series data without losing any potentially important detail. For this purpose, one can make use of a *goodness of fit* statistic which is based on the Euclidian Distance between the individual observations (i.e., time slots) and their appropriate cluster centres:[11] the smaller this distance on average (i.e., across all time slots), the better the fit. Of course, a large k (number of clusters) will always result in a better fit than a small k – if k were equal to the number of time slots, the fit would be perfect. We thus need to find a setting for k that provides an optimal compromise between goodness of fit on the one hand and a small number of clusters on the other.

The solution to this problem is to perform a series of cluster analyses with incrementally increasing settings of k, and to observe the resulting changes in fit. For the data in Figure C, the following pattern emerged (Figure E).

As can be seen, settings below $k = 12$ (highlighted with a black marker symbol) lead to quite dramatic impairments in fit (average distance abruptly increases with $k < 12$), whereas settings higher than $k = 12$ yield comparatively minor improvements (average distance gradually levels off). The interpretation of this graph is similar to that of *scree plots* in factor analysis. Hence, it can be concluded that $k = 12$ is the optimum setting for the data in Figure C.

A more confident estimate of the optimum k can be obtained by comparing these proportional increases in fit (i.e., decreases in average distance) against cluster analysis results from data that do not contain any structure at all except for the linear progression in time. The grey curve in Figure E shows the results from such a series of analyses: the corresponding data set was of the same size ($n = 150$, four conditions) as the time series in Figure C, but instead of actual data, each time slot contained random z-scores per condition. As can be seen, even for time series that contain nothing but random noise, there are proportional improvements in fit with increasing numbers of clusters. However, these improvements proceed far more smoothly

[10] Put differently, while the procedure maximises between-*cluster* differences (indeed, absolute effect sizes across the four clusters yield a standard deviation of 0.85 for the *k-means clustering* solution and of 0.60 for either of the alternative solutions), maximising between-*condition* differences per cluster is obviously not the same thing.

[11] The *k-means clustering* procedures in SPSS, Statistica, etc., provide this information on request.

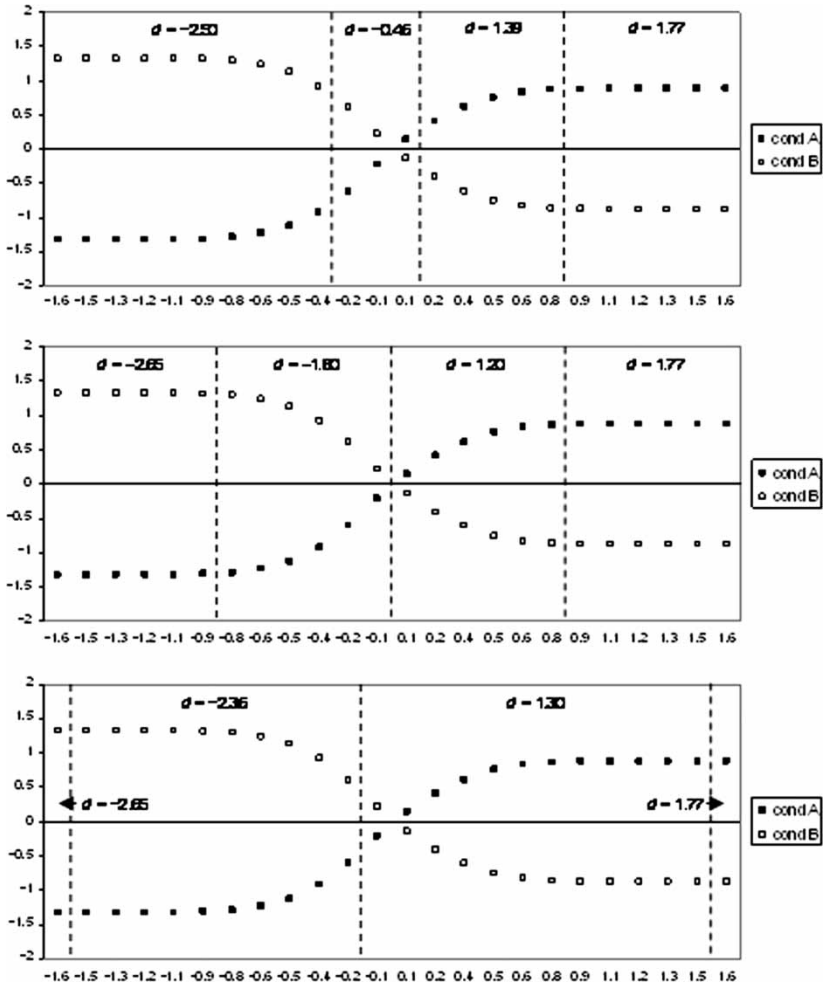

Figure D. Idealised data over 24 time slots and two conditions. The top panel shows a *k*-means clustering solution (dashed vertical lines; $k = 4$) using time and the relevant condition scores as clustering dimensions. The middle panel shows a four-cluster solution with equal spacing of clusters. The bottom panel shows a four-cluster solution in which the average absolute effect size per cluster is maximised. The d-values in each panel refer to mean cross-condition differences (A–B) over the time slots per cluster.

than for the actual data. Most importantly, note that the comparison between the two graphs corroborates the optimal setting of $k = 12$ for the 'real' data in Figure C: with $k > 12$, proportional improvements in fit largely follow the random data pattern.

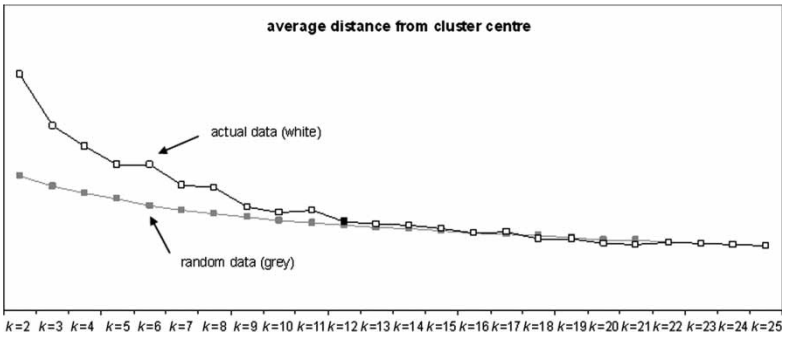

Figure E. Goodness of fit (average distance between individual observations and their relevant cluster centres) obtained from a series of cluster analyses varying k from 2 to 25. White marker symbols refer to analyses performed on the data in Figure C (the optimum number of clusters [$k = 12$] is highlighted by a black marker symbol). Grey marker symbols refer to results from a series of cluster analyses performed on random data (see text). For better comparison, goodness of fit is measured on a proportional scale.

Discussion

K-means cluster analysis is a useful tool for detecting structure in data. Combined with heuristics determining the optimum k (as described above), it preserves potentially important detail while at the same time being more objective and precise than a 'visual inspection' approach. However, one also needs to be aware of some major limitations of clustering algorithms. First of all, *k-means clustering* is not an inference-statistical method, and hence there is no guarantee that clustering results for a given data set will generalise to new samples of participants or items, respectively. Second, as with any other statistical procedure, *k-means clustering* results are largely dependent on data quality. For example, our own simulations indicated that high levels of random noise within the data set will increase the likelihood of temporally incoherent clustering results (i.e., clusters will be widely spread over time in a discontinuous manner). To counteract this problem, one could increase the weight of the time variable relative to the other clustering dimensions. However, there is only so much one can do about noisy input; collecting more data is certainly preferable. Third, there is no guarantee that *k-means clustering* will always come up with a theoretically relevant solution. The algorithm itself knows nothing about psycholinguistics; all it does is uncover structure in data. It is therefore perfectly legitimate to further adjust a given clustering solution in accordance with additional theoretical requirements. An example of this is our time window definition for Experiment 1. The clustering algorithm identified nine time windows with maximally distinct data patterns. However, one of these clusters spanned over a time period of -200 to $+400$ ms relative to the onset of the critical word. To be able to make strong assertions about anticipatory processing, we decided to split this cluster along the onset of the critical word itself. (Note that by sheer coincidence, the clustering solution for Experiment 2 was already 'aligned' with the onset of the critical word.) A final point concerns the sensitivity of this approach regarding fine-grained differences in processing dynamics. In this respect, all methods that are based on a reduction in temporal resolution (i.e., by aggregating time slots into larger analysis windows) have a clear disadvantage compared with more sophisticated curve-fitting approaches (e.g.,

Magnuson et al., 2007; Scheepers et al., 2008). However, the latter can be rather tedious and in the worst case even unfeasible: some time series data are simply too complex to be modelled in terms of a computationally (and theoretically) tractable function. Also, if one is primarily interested in whether a given effect starts before or after the onset of a critical word (i.e., without bothering too much about how *exactly* corresponding effect sizes develop over time), then data reduction methods can already reveal a sufficiently informative picture. Hence, as always, there is no single 'standard' method that would optimally apply to all kinds of statistical, theoretical, and practical problems one might encounter. Researchers have to choose among a range of possible solutions available, based on which method suits their problem best. *K-means clustering* is one such method, but in the end, the human brain remains the most important tool of all.

Experimental items used in Experiment 1

1.
If a plumber had the appropriate tools, he could do his job a lot faster.
If a plumber were trained in medicine, he would be very useful indeed.
Sarah could call a plumber to fix her broken [sink/foot] and she would be very relieved.

2.
If cats are hungry, they usually pester their owners until they get fed.
If cats were vegetarians, they would be cheaper for owners to look after.
Families could feed their cat a bowl of [fish/carrots] and listen to it purr happily.

3.
If Peter were looking for a cheap holiday, travel agents would have lots of suggestions.
If New York City were a city on the moon, it would be very expensive to get there.
Peter could fly to New York City in a [plane/ spaceship] and quietly admire the picturesque scenery.

4.
If sheep were very hungry, they are likely to help themselves to any food they find around them.
If sheep were carnivorous like wolves, they would be a lot less work to be looked after.
Farmers could leave their sheep in the field to eat [grass/rabbits] and concentrate on other farm work.

5.
If a tramp felt tired, he would have to look out for somewhere to rest for a while.
If a tramp won the lottery, he would be able to use the money to change his life.
The tramp could afford to live in a [shack/palace] and share it with cheerful friends.

6.
If hunters wanted to plan an event in the countryside, there would be lots of preparation involved.
If Wales had the same wildlife as Kenya, it would lead to lots of interest in the animals.
Hunters could go to Wales to hunt [foxes/elephants] and keep their activities a secret.

7.
If someone wanted a relaxing holiday abroad, I would recommend an island paradise like Hawaii.
If Hawaii had the same weather as the Arctic, the locals would need to keep warm in winter.
Holidaymakers could visit Hawaii to stay in a small [villa/igloo] and admire views from their window.

8.
If penguins want to survive, they must understand the risks in their home environment.
If all penguins lived in the African desert, they would have to adapt to the environment.
Penguins could learn to outrun polar bears/cheetah and find food for their young.

9.
If we had enough money for a holiday to Egypt, books would advise us where to visit.
If Egypt had the same landscape as Switzerland, it would be a beautiful country to visit.
Gary could go to Egypt to climb the [mountains/sand dunes] and get suntan to his face.

10.
If parents knew more about the causes of tooth decay, they would tailor the family's diet accordingly.
If tooth decay were only caused by eating vegetables, parents would tailor the family's diet accordingly.
Children could get tooth decay from eating too many [sweets/carrots] and cry about visiting the dentist.

11.
If a couple fancied some Italian cuisine, they are likely to choose a traditional dish.
If Italian cuisine were the same as Chinese, we would buy cookbooks for the recipes.
Lovers could go to an Italian restaurant to eat a [stir-fry/pizza] and feel very full all night.

12.
If a golfer were keen to win the tournament, he would need to sharpen up his technique.
If golf used the same equipment as tennis, players would work hard to improve their technique.
A golfer could hurt his shoulder from swinging his [club/racket] and visit a physio for treatment.

13.
If a caterpillar had eaten enough leaves, it would be ready to transform into its next stage.
If caterpillars turned into birds in the chrysalis, it would be an amazing transformation to study.
A caterpillar could mature into a [butterfly/sparrow] and enjoy testing its new wings.

14.
If you had planned to visit America, it is recommended that you watch a local sporting event.
If America had the same national sport as Spain, it would be a popular destination for enthusiasts.
Visitors could go to America to watch [baseball/bullfights] and join the crowds of spectators.

15.
If you monitored a spider's daily activities, you would be fascinated by their accomplishments.
If spiders had the same biological systems as bees, we would see evidence of their hard work.
A spider could spend all day producing [webs/honey] and admire its hard work later.

16.
If Mum wanted to impress her friends and relatives, she would have to work hard.
If margarine contained a detergent, it would have many useful domestic uses.
Mum could use margarine in her [baking/hair] and impress her friends and family.

17.
If travellers in France were interested in tasting local delicacies, they would enjoy the experience.

If France had the same cuisine as Japan, it would be popular with gastronomic enthusiasts.
They could go to France to eat [croissants/sushi] and recreate the recipes at home.

18.
If vets updated their training every few years, there would be lots of new techniques to learn.
If vets were only trained to treat mythical animals, it would be a complicated course to complete.
Vets could learn how to treat injured [puppies/unicorns] and write books to teach others.

19.
If building bricks in a house caught fire, it can cause substantial damage to a home.
If all building bricks were made from ice, safety would be an issue in their maintenance.
A house fire could cause the bricks to [burn/melt] and many firemen would be called.

20.
If you were interested in Australia's wildlife, a holiday 'down under' is well worth the money.
If Australia were inhabited solely by dinosaurs, expensive holiday tours would be popular.
Tourists could go to Australia to observe [kangaroos/Triceratops] and to buy special Aussie souvenirs.

21.
If Darwin had made more time for his research, he would have given us much more to learn from.
If Charles Darwin had been famous for his pharmaceutical work, we would learn about it in school.
Darwin could have published a book about the evolution of [animals/aspirin] and sold lots of copies.

22.
If you wanted to see polar bears in the wild, you would need to go to their natural environment.
If polar bears had evolved to favour hot climates, we would be interested in their behaviour.
We could see polar bears roaming the [Arctic/Rainforest] and observe their relations with other species.

23.
If a fast food joint were very quiet at lunchtimes, it would rely on customers' endorsements.
If a fast food joint only sold animal supplies, it would rely on customers' endorsements.
John could go to the fast food joint to buy some [chips/cat] food and recommend it to many friends.

24.
If meat were in short supply, workers would need to work hard to keep up with public demand.
If all meat were grown in test tubes, no animals would need to be killed to make sausages.
Pork meat could be produced in a [farm/lab] and we'd enjoy a good dinner.

Experimental items used in Experiment 2

1.
Janet unpacked the belongings and put the postcard in the cupboard.
[Later, Janet saw Barry move/ While Janet was busy, Barry moved] the postcard from the cupboard to the drawer.
Later, Janet wanted to see the postcard so she looked in the [cupboard/ drawer] and smiled.

2.
Mary planted all her seeds in trays and put them on the shelf.
[Later, Mary spotted her husband move/While Mary was distracted, her husband moved] the plants from the shelf to the table.
Later, Mary wanted to look at her plants so she looked on the [shelf/table] and daydreamed.

3.
Joanne collected her clothes and put them into the laundry basket.
[Later, Joanne saw Alex move/But while Joanne was unaware, Alex moved] the clothes from the laundry basket to the washing machine.
Later, Joanne wanted to check the clothes so she looked in the [laundry basket/washing machine] and sighed.

4.
John washed the dishes after his breakfast and left his watch on the table.
[Later, John noticed Victoria move/While John was dressing, Victoria moved] the watch from the table to the bed. Later, John wanted to find his watch so he looked on the [table/bed] and yawned.

5.
Mum bought the Christmas presents and hid them all in the wardrobe.
[Later, Mum watched Dad move/While Mum was busy, Dad moved] all the presents from the wardrobe to the antique chest.
Later, Mum wanted to see the presents again so she looked in the [wardrobe/antique chest] and grinned.

6.
Kevin returned from the shops and put his chocolate in the fridge.
[Later, Kevin noticed Jessica move/While Kevin was out, Jessica moved] the chocolate from the fridge to her handbag. Later, Kevin wanted to eat the chocolate so he looked in the [fridge/handbag] and squealed.

7.
Margaret washed her favourite shirt and put it away in the drawer.
[Later, Margaret spotted Russell move/While Margaret was at work, Russell moved] the shirt from the drawer to the basket.
Later, Margaret wanted to find the shirt so she looked in the [drawer/basket] and paused.

8.
Julie entered the classroom and put her finished homework in her desk drawer.
[Later, Julie spotted Maxine move/While Julie was distracted, Maxine jokingly moved] the homework from the drawer to her bag.
Later, Julie wanted to check her homework so she looked in the [drawer/bag] and focused.

9.
Maria packed her suitcase to go on holiday and put the tickets in her suitcase.
[Later, Maria saw her dad move/While Maria was showering, her dad moved] the tickets from the suitcase to her purse.
Later, Maria wanted to check the tickets so she looked in the [suitcase/purse] and giggled.

10.
Colin bought a big bottle of vodka and put it in the drinks cabinet.
[Later, Colin noticed Angela move/While Colin was unaware, Angela moved] the whisky from the drinks cabinet to the freezer.
Later, Colin wanted to drink the whisky so he looked in the drinks [cabinet/freezer] and moaned.

11.
Lauren wrote about her day in her diary every night and always kept it in a box.
[One day, Lauren spied her brother move/One day, Lauren's brother secretly moved] the diary from the box to the wardrobe.
Later, Lauren wanted to find her diary so she looked in the [box/wardrobe] and groaned.

12.
Jamie celebrated buying the house by putting a small tree in the veranda.
[Later, Jamie saw his dog dig up the tree and drag/While Jamie was inside, his dog dug up the tree and dragged] it from the veranda to the kennel.
Later, Jamie wanted to see the tree so he looked in the [veranda/kennel] and frowned.

13.
Linda arrived at work and put her briefcase on the desk.
[Later, Linda noticed Alan move/While Linda was unaware, Alan moved] the briefcase from the desk to the chair.
Later, Linda wanted to check her briefcase so she looked on the [desk/chair] and groaned.

14.
Laura left the hostel for breakfast and left her rucksack in the bed.
[Later, Laura saw the cleaner move/While Laura was out, the cleaner moved] the rucksack from the bed to the locker.
Later, Laura wanted to collect her rucksack so she looked in the [bed/locker] and gulped.

15.
Gillian cooked a casserole and left it to cool down in the oven.
[Later, Gillian spotted Mark move/While Gillian was not looking, Mark moved] the casserole from the oven to the fridge.
Later, Gillian wanted to eat the casserole so she looked in the [oven/fridge] and salivated.

16.
Tony brought his concert tickets into work and put them in his coat pocket.
[Later, Tony spotted Gary move/While Tony was in a meeting, Gary moved] the tickets from the pocket to the bin.
Later, Tony wanted to examine the tickets so he looked in the [pocket/bin] and coughed.

17.
Mum finished making desserts for the birthday party and put them in the oven.
[Later, Mum watched Dennis move/While Mum was showering, Dennis moved] the cakes from the oven to the cupboard.
Later, Mum wanted to taste a cake so she looked in the [oven/cupboard] and drooled.

18.
Max bought beer for his flat party and put it in the lounge.
[Later, Max saw Charlie move/While Max was not looking, Charlie moved] the beer from the cupboard to the fridge.
Later, Max wanted a beer so he looked in the [cupboard/fridge] and frowned.

19.
Shane finished his vigorous gym workout and left his trainers in the locker.
[Later, Shane noticed the gym staff move/While Shane was at a lecture, the gym staff moved] the trainers from the locker to the bin.
Later, Shane wanted his trainers so he looked in the [locker/bin] and shivered.

20.
Dennis bought the expensive Christening present and hid it in the safe.
[Later, Dennis watched Chloe move/While Dennis was gardening, Chloe moved] the present from the safe to her handbag.
Later, Dennis wanted to check the present so he looked in the [safe/handbag] and smiled.

21.
Janis arrived at work and put the newspaper article in the filing cabinet.
[Later, Janis noticed Stephen move/But while Janis was occupied Stephen moved] the article from the filing cabinet to the desk.
Later, Janis wanted to collect the article so she looked in the [filing cabinet/desk] and tutted.

22.
Doug looked at the beautiful wedding photo and put it on the fireplace.
[Later, Doug saw Mandy move/While Doug was gardening, Mandy moved] the album from the fireplace to the bookcase.
Later, Doug wanted to see the album again so he looked on the [fireplace/bookcase] and daydreamed.

23.
Sophia flicked through the pharmacy journal then put it on the chair.
[Later, Sophia saw the librarian move/While Sophia was distracted, the librarian moved] the journal from the chair to the shelf.
Later, Sophia wanted to read the journal so she looked on the [chair/shelf] and whistled.

24.
Isobel entered the beautician's room and put her jewellery on the bed.
[Later, Isobel spotted the therapist move/Without Isobel's knowledge, the therapist moved] the jewellery from the bed to the chair.
Later, Isobel wanted to put her jewellery back on so she looked on the bed/chair and relaxed.