Published in final edited form as:

J Exp Psychol Learn Mem Cogn. 2011 May ; 37(3): 748–756. doi:10.1037/a0022637.

Reversing the hands of time: Changing the mapping from seeing to saying

Stefanie E. Kuchinsky,

Medical University of South Carolina

Kathryn Bock, and

University of Illinois

David E. Irwin

University of Illinois

Abstract

To describe a scene, speakers must map visual information to a linguistic plan. Eye movements capture features of this linkage in a tendency for speakers to fixate referents just before they are mentioned. The current experiment examined whether and how this pattern changes when speakers create atypical mappings. Eye movements were monitored as participants told the time from analog clocks. Half of the participants did this in the usual manner. For the other participants, the denotations of the clock hands were reversed, making the big hand the hour and the little hand the minute. Eye movements revealed that it was not the visual features or configuration of the hands that determined gaze patterns, but rather top-down control from upcoming referring expressions. Differences in eye-voice spans further suggested a process in which scene elements are relationally structured before a linguistic plan is executed. This provides evidence for structural rather than lexical incrementality in planning and supports a “seeing-for-saying” hypothesis in which the visual system is harnessed to the linguistic demands of an upcoming utterance.

Keywords

Language Production; Eyetracking; Time-telling

Common expressions have conventional word orders. People prefer to say *bread and butter* over *butter and bread* or *red, white, and blue* over any other sequence (e.g., Cooper & Ross, 1975; Kelly, Bock, & Keil, 1986; Onishi, Murphy & Bock, 2008). Similarly, in contemporary American English, *one-forty* is favored over *twenty till two* when time is told from an analog or digital clock (Bock, Irwin, Davidson, & Levelt, 2003). This highly ingrained ordering is associated with a highly ingrained way of extracting information from clocks. Speakers almost always look to the hour hand and then to the minute hand just before talking (Bock et al., 2003; Bock, Irwin, & Davidson, 2004; Davidson, Bock, & Irwin, 2003). Even when instructed to produce a disfavored structure in which minutes precede

Correspondence: Stefanie E. Kuchinsky, Medical University of South Carolina, Department of Otolaryngology, 135 Rutledge Ave, MSC 550, Charleston, SC 29425, kuchins@musc.edu.

Publisher's Disclaimer: The following manuscript is the final accepted manuscript. It has not been subjected to the final copyediting, fact-checking, and proofreading required for formal publication. It is not the definitive, publisher-authenticated version. The American Psychological Association and its Council of Editors disclaim any responsibility or liabilities for errors or omissions of this manuscript version, any version derived from this manuscript by NIH, or other third parties. The published version is available at www.apa.org/pubs/journals/xlm

hours (as in *twenty till two*), speakers may continue to look at the hour hand earlier and more often than the minute (Bock et al., 2003), changing their fixation patterns only as they adapt to the unfamiliar expressions (Kuchinsky, Bock, & Irwin, 2006). In this experiment, we evaluated whether and how these visual preferences drive or are driven by linguistic preferences.

The normal correlation between where people look and what they say makes it hard to untangle the influence of visual and conceptual factors on language use. The problem mirrors the longstanding debate about the roles of bottom-up and top-down influences on vision. Perceptual properties such as color, visual complexity, luminance, and contour affect attention and eye movements to parts of scenes (see Irwin, 2004 for review). However, at least since Yarbus (1967), it has been known that early eye-movement patterns reflect such top-down factors as viewers' goals. These effects can arise rapidly: Viewers may extract the gist of a scene within 100ms (Potter, 1976) and use the information to locate informative areas (see Henderson & Ferreira, 2004 for review). The question addressed here is what the eyes do when the goal is to talk about what one sees. That is, when the goal is to create and execute a language plan, what does the visual system do?

There are two general hypotheses about the link between looking and talking, each of which relates a perceptual process to a language process. One emphasizes bottom-up control from perception. The *perceptual guidance* hypothesis predicts that initial targets of attention drive word order in language (e.g., Flores d'Arcais, 1975; Osgood, 1971; Osgood & Bock, 1977; Tomlin, 1997). The most convincing support for perceptual guidance comes from Gleitman, January, Nappa, and Trueswell (2007), who used a subtle attention-capture manipulation to cue where one of two actors would appear in a scene. Speakers were more likely to start their scene descriptions with a cued actor than an uncued one, consistent with a model of language production in which visual information immediately activates lexical information, which is then likely to be used to initiate an utterance (Bock, 1982).

The intuition that there is a natural, causal linkage between the initial focus of attention and the beginnings of utterances is so compelling that it survives despite relatively weak evidence (see Bock et al., 2004) and in the face of paradigmatic shifts in psychological theory. It was central to behaviorist and neobehaviorist accounts of sentence production (Ertel, 1977; Lounsbury, 1965; Osgood, 1971; Osgood & Bock, 1977) and remains popular in more contemporary cognitive perspectives (Brown-Schmidt & Konopka, 2008; Flores d'Arcais, 1975; Gleitman et al., 2007; Levelt, 1981; MacWhinney, 1977; Tomlin, 1995, 1997; van Nice & Dietrich, 2003). According to this view, language plans unfold naturally from perceptually controlled starting-points. The implication is that people may begin their utterances without having any idea about what they will say next. Sentences can be constructed on the fly by sequentially associating concepts in the order in which they become perceptually or conceptually most salient (Paul, 1886/1970).

The hypothesized coupling between an immediate and potent attentional influence on upcoming utterances and the incremental projection of concepts or words posits a mapping mechanism that we call *lexical incrementality*. Generally, lexical incrementality is fundamental to language production accounts in which lexical retrieval drives the creation of a sentence's structure (e.g., Branigan, Pickering, & Tanaka, 2008; Kempen & Hoenkamp, 1987; de Smedt, 1996; van Nice & Dietrich, 2003). It suggests that preparation time for a word or constituent in a sentence is chiefly a function of properties like its complexity or frequency (Meyer, 2004). When an object is viewed, its name may be activated, increasing the likelihood that the word will be phonologically encoded (Gleitman et al., 2007). Other visual elements influence language planning during later, successive fixations. Accordingly,

there is a natural interplay between extracting visual information, preparing a corresponding word or phrase, and then repeating the cycle until an utterance is complete.

An alternative to perceptual guidance with lexical incrementality involves *structural guidance* with structural incrementality (Bock et al., 2003). Structural guidance emphasizes the wholistic uptake of information and its role in creating a top-down language plan that channels an analysis of visual information. This *seeing-for-saying* hypothesis is perceptually akin to “thinking for speaking” (Slobin, 1996), and assumes that a relational or causal structure can emerge from glimpses of events. This structure serves as the starting point for an incomplete utterance framework that can exert top-down control of how gazes are directed toward elements of a scene. On this kind of account, a rudimentary syntactic plan guides successive eye movements to the elements of an event. In short, a provisional plan for what is to be said drives the pattern of looking.

The mapping mechanism that follows from structural guidance involves what we call *structural incrementality*. Structural incrementality entails the creation of a linguistic action plan that provides information about where to start and how to continue an utterance. Though the plan plays out in terms of a sequence of lexical events, an overarching structure controls the sequence. Even with lexical-syntactic interaction (F. Ferreira & Engelhart, 2006; V. Ferreira & Dell, 2000), the generation of sentence structure precedes filling of lexical slots, even when lexical availability guides the process.

Griffin and Bock (2000) provided some of the first evidence for structural-guidance. They presented participants with pictured events that could be naturally described with either active or passive structures, and compared the eye movements that occurred to the pictures when participants performed linguistic or nonlinguistic tasks. Some participants described the events and others simply inspected the pictures in anticipation of judging aesthetic quality. A third group identified the undergoer of an action (i.e., the patient), a task whose successful performance depends on apprehending an event’s causal structure.

Simple picture inspection served to gauge the impact of relative perceptual salience on first looks to the events. The initial fixation targets during inspection were in fact not the elements that were typically fixated first on the event description trials, which normally became the sentence subjects. This suggests that general salience did not drive first mention or subject selection.

The comparison between event description and causal construal (patient identification) suggested a different source for control of subject selection. In the causal construal task, an early, brief interval of event viewing was followed by eye movements to the patient, suggesting that initial viewing served to extract the event’s causal structure (roughly, who did what to whom). In event description, there was a similarly brief event-viewing interval, followed directly by an eye movement to the element of the scene that would serve as the sentence subject. Evidently, the first glimpse provided enough information to build the beginnings of an appropriate sentence structure. Like patient identification, this process depends on an apprehension of event structure that can support the construction of a simple sentence. What ensues is a fast and fluent unfolding of an utterance in which time-locked looks to scene elements precede corresponding phrases. The structural plan provides enough information to send the eyes to relevant locations, with *name-related gazes* (Griffin, 2004) supporting the sequencing and detailing that fills the structure out.

More evidence for structural incrementality comes from experiments on telling time from analog and digital clocks. Bock et al. (2003) found that when speakers produced absolute (e.g., *one-forty*) or relative (e.g., *twenty till two*) time expressions, the latencies from fixating an object to producing its name were longer for the first than for the second term of time

expressions. Most revealing was that these *eye-voice spans* were affected by compatibility between the clock display and the linguistic form of the expression, but only for the first term. For digital displays, compatibility arises in the dovetailing of a left-to-right spatial array of time information with the form of absolute expressions (1:40 is *one-forty*). For analog displays, the necessary parsing is relational, which is more compatible with relative expressions (i.e., the angle of the hands more readily conveys the relation between twenty and two in *twenty till two* than digital displays, which contain neither a twenty nor a two). Thus, when the structure that links the first term and the second is supported by visual properties, projection beyond the first phrase of a structure is easier than when a scene affords no simple scaffolding for the connection between the first and second term. So, the modulation of first- and second-phrase preparation by perceptual-linguistic compatibility suggests structural incrementality in linguistic mapping.

Our goal in the present experiment was to use time-telling to test perceptual guidance (encompassing lexical incrementality) against structural guidance (encompassing structural incrementality). We used analog clock displays to elicit time expressions and varied the compatibility between the hand configurations and time expressions by creating normal or inverted time-telling tasks (Figure 1).

In the normal time-telling task, the mapping from clock hands to absolute time expressions works conventionally: The little hand denotes the hour and the big hand denotes the minute. In the absolute time expressions spontaneously used by our participants, the hour is named first, using the term for the number position to which the hand actually or virtually points (*eleven* in Figure 1). The minute is named second, using what is mathematically a multiple of the number position to which the hand points (e.g. in Figure 1, *forty* is five times the standard position number, which is *eight*). So, the little hand on the eleven and the big hand on the eight is *eleven-forty*. In effect, during normal time-telling there is a transformation in the numbering framework from a big-hand schema in which the position-names range from *one* through *12*, to a little-hand schema in which the names range from *five* through *fifty-five*. (We do not assume that multiplication actually occurs in this transformation, only a change in the numbering framework in which naming occurs.)

In the inverted task, the order of mentioning the hands changes and the link between the hands and the numbering frameworks is reversed. The big hand (normally the minute hand) must be named first, in terms of the hour number framework, while the little hand (normally the hour hand) is named second, in terms of the minute framework. Thus, the reversal in the mapping from big and little hands to the appropriate names requires a change in the habitual relationship between vision and language in time-telling.

Given this task analysis, perceptual guidance suggests that production of time expressions arises in an inescapable lifelong association between smaller clock hand and the starting point for a time expression (i.e., the hour name). In laboratory tasks, *automatic detection* in visual search develops when there is a consistent mapping between a perceptual target and a response. Schneider and Shiffrin (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) found that with consistent mapping between a target category (e.g., numbers or letters) within a set of distractors, participants learned to detect targets effortlessly, with no processing costs for additional items. The learned association also resisted modification with verbal instructions. In real-world tasks, too, extensive practice leads to automatic detection of relevant stimuli including everything from airline luggage screening (Madhavan & Gonzalez, 2006) to chicken sexing (Biederman & Shiffrin, 1987). Adult time-tellers have years of experience with consistent mapping from the smaller hand to the first time expression (the hour term) and from the bigger hand to the second time expression (the minute term). According to the perceptual guidance account, attention (and the eyes) should

be automatically attracted to the little hand by force of habit, accounting for the tendency for speakers to produce hour terms before minute terms during time-telling. Consequently, perceptual guidance predicts that switching the hands so that small hands denote minutes and big hands denote hours, violating a lifetime of consistent-mapping experience, should disrupt utterance formulation.

The predictions from structural guidance are different. When a spoken utterance is the goal of scene viewing, top-down control from a language plan may allow a more flexible uptake of visual information. The implication is that the nature of the expression rather than the properties of the scene should control eye-movements. From this perspective, any tendency to seek the little hand first stems from a preference to name hours before minutes. Finding the little hand is not the immutable product of perceptual detection, but of the goal to tell the time in a customary way. If telling the time in a customary way demands a different parse of the scene, changing the names of the clock hands should lead to a fast, fluent inversion of the order in which the big and little hands are fixated.

The perceptual and structural guidance hypotheses make additional predictions that follow from the kinds of language plans that they naturally support, respectively lexical and structural incrementality. Both kinds of incrementality imply that finding the hour hand should be followed by producing the hour expression, and finding the minute hand should be followed by producing the minute expression. However, lexical and structural incrementality predict different temporal patterns in the relationship between eye movements and the elements of time expressions.

Specifically, lexical incrementality predicts that eye-voice spans for the first and second terms should be equal for the hands in both types of time-telling. That is, once the appropriate hand is fixated, the eye-voice spans for each component of the time expressions should be predictable from properties of the individual expressions themselves, like linguistic complexity (e.g., “oh-five” is more complex than “five”). So, with inverted instructions, longer eye-voice spans should accompany both terms of the time expression, protracted to roughly equal degrees relative to regular time-telling. In contrast, structural incrementality implies the readying of more initial structure. When telling time from an analog clock, the structural foundation for an absolute expression requires a transition between the naming frameworks for the hands. If the transition between these frameworks is prepared in advance of expression onset, initiation of the first term should be more delayed relative to initiation of the second term in inverted than in normal time-telling.

In the current experiment, half of the speakers told time from analog clocks in the normal manner. The other speakers had to invert the normal naming frameworks of the hands, treating the big hand as indicating the hour and the little hand as indicating the minute. Participants received no instructions about what time expressions to use, with the expectation that they would spontaneously produce absolute expressions like *one-forty* (as they did). Participants also received no instructions about how to view the clock, leaving the process of extracting information from the scene to natural mechanisms and spontaneous adjustments. In short, the only instructions involved how the clock hands related to hour and minute designations.

To summarize, the perceptual guidance hypothesis predicts that participants should look first to the little hand and should persist in doing this regardless of whether the hand is taken to name the hour or the minute. Structural guidance instead predicts an immediate shift to big-hand first fixations in the inverted condition. Predictions about eye-voice spans follow from the types of language preparation that follow naturally from perceptual and structural guidance. Perceptual guidance predicts that the time from fixations to utterance terms should

increase roughly equivalently for the first and second terms under inverted time-telling, reflecting lexical incrementality. Structural guidance predicts an increase in the first-term/second-term difference under inverted time-telling, reflecting structural incrementality.

Method

Participants

Sixteen undergraduate students participated in this experiment for credit in an introductory psychology class. All were native English speakers who reported normal or corrected-to-normal hearing and vision.

Materials and Design

The stimuli were black-on-white line drawings of analog clocks (Figure 1). The clocks lacked tick marks and numbers, which does not significantly affect the overall pattern of clock viewing or time-telling accuracy (Davidson et al., 2003). The times displayed included all 144 times that are multiples of five minutes on a 12-hour clock. The little hand pointed directly at one five-minute location (i.e. at 30° intervals from :00) regardless of where the big hand was located. This was done to equate the movement of the hands under both regular and inverted time-telling. For example, in Figure 1, one hand points directly at the eight and the other at the eleven, allowing the responses *eleven-forty* and *eight fifty-five* under normal and inverted instructions respectively. (On an ordinary analog clock, the little hand transits proportionally from one hour to the next as the big hand transits from minute to minute. To compare displays of this kind to the discrete displays used, a separate time-telling condition was run in which participants told time normally from clocks in which the hour hand varied continuously between successive hours, in line with minute-hand changes. There were no significant differences in the results for the two types of hour movements across voice onset measures and all hour and minute gaze onset and duration times.)

Eight stimulus presentation lists were created with different pseudo-randomizations of all 144 clock times. Consecutive trials with the same hour or minute positions were prohibited. Each list was presented to two participants, one participant under regular and one under inverted instructions, so that each list occurred once in each of the two conditions.

Apparatus

Eye movements were recorded with a headband-mounted Eyelink II eyetracker (SR Research, Ltd.) controlled by a Dell Optiplex GX270 computer. The eyetracker had 500Hz temporal resolution and 0.2° spatial resolution. The stimuli were displayed on a 21" SVGA color monitor with a screen resolution of 800 x 600 pixels and a refresh rate of 85Hz. Verbal responses were converted to digital sound files via a Creative Labs SoundBlaster X-Fi XtremeGame sound card over a preamplified 500-ohm unidirectional dynamic microphone. A hand-held Eyelink button box recorded button presses.

Procedure

Clocks were centered on a computer monitor with the clock face circumscribing a circle of approximately 25.4 cm diameter. Participants viewed the screen from a distance of about 49 cm, such that the clocks subtended a visual angle of approximately 29°.

Before every trial, a drift-correct phase required that participants fixate the center of the screen. An analog clock was then displayed for three seconds. Half of the participants were instructed to tell the time aloud from the clock as they normally would. An example trial with a display like the one in Figure 1 was provided, along with a corresponding expression (e.g., *eleven-forty*). The other half of the participants were instructed to tell the time in

another way, treating the little hand as the minute and the big hand as the hour, with an example like Figure 1 described as showing *eight fifty-five*. (The example probably increased the use of absolute expressions, but the use of such examples served only to increase production from a level near 90% to almost 100% in Bock et al., 2004.)

All participants were asked to avoid producing disfluencies and to tell the time as quickly and accurately as possible. After telling the time, participants pressed a button to proceed to the next trial. Instructions were followed by four practice trials in which the experimenter corrected any mistakes the participants made and provided more examples if needed.

Scoring

There were 12 trials per list on which the hour and minute hands overlapped, leaving only one hand visible (e.g., *eleven-fifty-five*). These were not scored, eliminating 96 trials out of 1152 possible trials per group, or 8% of the data. Trials were also excluded when the participant was disfluent, saying “uh” or “um” or self-correcting (11%) and when the voice onset time was less than 300ms or more than 4000ms (7%). Finally, trials on which participants never fixated the hour and minute hands were excluded (14%). An average of 39% of the trials were lost for each participant (41% for normal time-tellers and 38% for inverted) with a range of 29–56%. The exclusions yielded 1401 analyzable trials, 49% regular and 51% inverted.

For analyzing fixation locations, each analog display was divided into twelve 30° slices that transected the clock face along its radii, bounding the five-minute points on the clock edge. For example, the boundaries of the *oh-five* minute region were the radii at 2 minutes 30 seconds (15°) and 7 minutes 30 seconds (45°). A saccade was defined as a change in eye position that covered more than .15° at a rate of at least 30° of visual angle/second.

To determine the onsets of hour and minute expressions in spoken responses, transcriptions were aligned with the sound files using the Sphinx-II speech recognition system and Viterbi forced alignment (see Bock et al., 2003, for detailed description).

Results

Accuracy and Voice Onset Times

All participants produced expressions with the hour term before the minute (e.g., *eleven-forty*). Participants were more accurate telling the time under inverted instructions ($M = 84.8\%$ correct) than regular instructions ($M = 72.0\%$), but this difference fell short of significance by a two-tailed t -test, $t(14) = 1.73$, $p = .11$. Participants were somewhat faster to begin telling the time (onset of the hour expression) under regular ($M = 1626\text{ms}$) than inverted instructions ($M = 1887\text{ms}$). This difference was also not significant, two-tailed $t(14) = 1.63$, $p = .13$. The inverse relationship between speed and accuracy suggests a weak tradeoff, but because our interest centered on the relationship between eye movements and speech onset, it does not compromise the results.

Onset times for the minute expression did not differ significantly between the regular ($M = 2075$) and inverted ($M = 2265$) time-telling groups, $t(14) = 1.48$, $p = .16$. Because this contrast on its own is of little importance, we do not discuss it further.

Eye Movements

Guidance—In order to evaluate the perceptual and structural guidance hypotheses, we report gaze patterns in two ways. When the visual form of the hand is at issue, we refer to looks to the “little hand” (in Figure 1 pointing to the eleven) and “big hand” (pointing to the

eight). When the name used to describe the hand is in question, we refer to looks to the “hour” and “minute.” For regular time-telling from the clock in Figure 1, the *eleven* position represents the hour referent and the *forty* position represents the minute referent. For the inverted group, the hour is *eight* and the minute is *fifty-five*. A gaze was defined as the sum of successive fixations on a hand without looks to any other part of the display.

Figure 2 depicts the pattern of eye movements. The dependent measures were mean gaze onset time, represented by the left edges of the bars, and gaze duration times, represented by the lengths of the bars, to the little and big hands. These measures are graphed against elapsed time, represented as a percentage of the interval between display onset and voice onset. These measures normalize over individual variations in latency to initiate an utterance, allowing for a more accurate comparison. The raw difference in voice onsets between groups was 261ms, which was non-significant.

The perceptual guidance hypothesis predicts that lifelong experience using the little hand to denote hours should drive initial gazes to the little hand in both instruction conditions. However, if a change in instructions readily overrides this habitual gaze pattern, with expression structure guiding looking, initial gazes should target the first-mentioned item, the little hand in regular time-telling and the big hand in inverted time-telling. Figure 2a shows that for regular time-telling, participants initially looked at the little hand (the hour hand) beginning on average after 22% of the interval before voice onset elapsed and spent an additional 44% of the interval gazing at it. They began looking at the big hand (minute hand) after about half of the voice-onset interval (46%) and spent 20% of the interval gazing at it. The 44% to 20% difference indicates that during regular time-telling, participants spent much more of their pre-speaking time gazing at the little hand than at the big hand, $t(7) = 10.44$, $p < .001$ by a paired two-tailed t-test.

Critically supporting structural guidance, the inverted group showed the same pattern in reverse (Figure 2b). Participants began looking at the *big* hand (the hour) after 19% of the voice-onset interval elapsed and spent 37% of the time gazing at it. They looked at the little hand (the minute) about halfway into the interval before voice onset (45% of voice-onset time) and spent about 20% of the time gazing at it. Participants spent much more pre-speaking time gazing at the big hand compared to the little hand, $t(7) = 4.90$, $p < .002$.

Figure 2c compares looking times to the clock hands during the voice-onset intervals for the two instruction types, with regular time-telling represented by solid bars and inverted time-telling by striped bars. For looks to the little hand, the 23% difference between regular and inverted instructions for mean gaze onset was significant, $t(14) = 7.43$, $p < .001$. The 24% difference in durations was also significant, $t(14) = 9.22$, $p < .001$. For looks to the big hand, there was a 27% difference between the groups in mean gaze onset, $t(14) = 14.00$, $p < .001$, and a 17% difference in duration, $t(14) = 5.19$, $p < .001$. So, the pattern of eye movements for the inverted clocks approached the inverse of the pattern for regular clocks.

In terms of hour and minute looking patterns, the regular and inverted groups were essentially the same. Speakers gazed at the hour hand at a similar point within the latency-to-voice-onset window (a 3% difference between regular and inverted time-telling) and for about the same duration (a 7% difference between regular and inverted instructions). Similarly, gazes to the minute hand occurred at about the same point (1% difference) for the same duration (0% difference).

Table 1 shows actual times for gaze onset, duration, and voice onset. The significance of comparisons did not change in analyses of the raw data.

Incrementality—Eye-voice spans were calculated to evaluate the hypotheses of structural and lexical incrementality. The eye-voice span represents the latency between fixating a referent and producing its name. For the hour hand in regular time-telling, the eye-voice span is the time between looking at the little hand and naming the hour; for inverted time-telling, the same eye-voice span is the latency between looking at the big hand and naming the hour. Analogous eye-voice spans were calculated for the minute hands in each condition.

The results for eye-voice spans are shown in Figure 3. Consistent with the predictions from structural incrementality, the mean eye-voice spans for term one (the hour) increased 263ms from regular to consistent mapping (1268ms vs. 1531ms, respectively, while the span for term 2 (the minute) changed only 94ms (from 1288ms to 1382ms, respectively).. In planned comparisons, the term-1 difference was significant ($t(14) = 2.23, p < .05$) and the term-2 difference was not ($t(14) = 1.25, p > .05$). The interaction between instruction type and expression position bordered on significance, $F(1, 14) = 4.38, p = .055$.

The term-1/term-2 eye-voice relationship persisted across the experiment for both regular and inverted time-telling, without substantial changes in gaze onsets or durations from the first to the second half of the experiment. There was a small general decrease in hour gaze-onsets in both conditions (32ms for regular time-telling and 29ms for inverted time-telling). Gaze durations for the hour increased by 30ms and 25ms for regular and inverted time-telling, respectively. Eye-voice spans and voice-onset times both decreased approximately 200 ms in the course of the experiment, at similar rates in both conditions, suggesting general practice effects.

Discussion

In this experiment, we compared perceptual and structural guidance accounts of how utterances are initiated and how the mapping from perception to language proceeds. We examined how participants told time from analog clocks under normal conditions (little hand designating hour and big hand designating minute) or under conditions that reversed the conventional relationship between clock displays and time expressions (big hand designating hour and little hand designating minute). The perceptual guidance hypothesis emphasizes the importance of perceptual factors in determining how speakers start utterances. For time expressions, years of perceptual learning should predispose speakers to look at the little hand first, regardless of how it is named. In contrast, structural guidance (“seeing-for-saying”; Bock et al., 2003) emphasizes the importance of a rudimentary utterance framework in determining visual starting points. Structural guidance predicts that first looks will target the hour hand, regardless of which hand (little or big) denotes the hour.

The results supported structural guidance: In both conditions, speakers initially attended to the hand whose referent occurred first in the utterance plan. With regular instructions, they looked first to the little hand; with inverted instructions, they looked first to the big hand.

The mapping mechanisms that follow from perceptual and structural guidance were evaluated with eye-voice span measures. Perceptual guidance with lexical incrementality implies that speakers plan utterances word by word, as each element is viewed. This predicts that eye-voice spans for term-1 and term-2 referents should be equally lengthened by the increased difficulty of the mapping under inverted conditions, as speakers instantiate the hour- and minute-naming frameworks separately. Structural guidance with structural incrementality predicts that speakers, early on, project abstract partial structures for utterances. Since inversion requires a novel, reversed naming framework at the outset of an utterance, the major change under inverted instructions should occur in the term-1 eye-voice span.

Consistent with structural incrementality, eye-voice spans under inverted time-telling were longer for term one than for term two, relative to regular time-telling. This finding argues that speakers did not re-create a vision-to-language mapping for every term of the expression. Instead, speakers appeared to prepare a structural mapping early in utterance planning, projecting a structure appropriate for upcoming words.

These outcomes serve to support an elaborated view of seeing-for-saying that addresses long-standing issues about utterance initiation (MacWhinney, 1977) along with long-standing issues about utterance preparation (Lashley, 1951). Going beyond Bock et al.'s (2003) work, the experiment tests a structural guidance view that combines predictions about how vision and language are coordinated in creating the starting points for connected speech with predictions about the mapping behind the sequential structure of utterances. The present research also increases the experimental rigor of the Bock et al. test, controlling for the nature of the displays (all of them were analog clocks) and the nature of the utterances (all of them were absolute expressions), varying only the degree of incompatibility between the visual display and the accompanying language.

The results do not mean that the visual properties of displays never influence word order or that utterances are never planned with lexical incrementation. A number of studies have shown that in some circumstances, speakers do plan word-by-word or simple phrase by simple phrase (e.g., Brown-Schmidt & Konopka, 2008; Brown-Schmidt & Tanenhaus, 2006; Gleitman et al., 2007). When scenes have little natural or inherent structure, or when speakers lack a coherent message to convey, it makes sense to begin with something salient and proceed linearly, one thing at a time. Indeed, Brown-Schmidt and Konopka (2008) suggest that their evidence for lexical incrementality in producing modified noun phrases (e.g., *the little house*) may represent a lower bound of message planning. In other circumstances, when the link from a scene or event to a linguistic expression is so often used as to be automatically instantiated, there may be so little need for preparation that structural incrementality leaves no trace. Regular time-telling in the present experiment approximates this state of affairs.

However, under the wide range of conditions in which events and scenes are described or ideas and thoughts are conveyed, there is no simple one-to-one mapping from vision to language. For describing events and scenes, figure-ground relationships have to be evaluated and resolved in order to separate what is worth mentioning from what is not. For ideas and thoughts, the point behind a particular message has to be untangled or decontextualized from a welter of cognitive detail, and linguistically conceptualized. All of these processes constitute a kind of groundbreaking (Bock et al., 2003), an initiation of production by speakers that is the inverse of the achievement of understanding by comprehenders. Speakers must distill or *disintegrate* what they intend to say from what they know; comprehenders must assimilate and integrate the contents of an utterance with existing knowledge.

In summary, we offer evidence that when speakers create mappings between vision and language, they can project the rudiments of utterance frameworks at the outset of linguistic formulation. The frameworks serve as action plans that guide changes in attention as details are spelled out, as speakers elaborate the frameworks with structural incrementation. We propose that utterance initiation can be driven not only by what draws visual attention, but by a linguistic plan to which visual attention is tethered.

Acknowledgments

The experiments and the preparation of the manuscript were supported in part by research and training grants from the National Science Foundation (BCS-0214270, BCS-0132292) and the National Institutes of Health (R01

MH66089; T32 MH18990). We thank Sarah Brown-Schmidt, Douglas Davidson, Brendon Hsieh, and Matthew Rambert for many helpful discussions and assistance.

References

- Biederman I, Shiffrar MM. Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual-learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 1987; 13:640–645.
- Bock JK. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*. 1982; 89:1–47.
- Bock K, Irwin DE, Davidson DJ, Levelt WJM. Minding the clock. *Journal of Memory and Language*. 2003; 48:653–685.
- Bock, K.; Irwin, DE.; Davidson, DJ. Putting first things first. In: Ferreira, F.; Henderson, J., editors. *The integration of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press; 2004. p. 249-278.
- Branigan HP, Pickering MJ, Tanaka M. Contributions of animacy to grammatical function assignment and word order during production. *Lingua*. 2008; 118:172–189.
- Brown-Schmidt S, Konopka AE. Little houses and casas pequeñas: Message formulation and syntactic form in unscripted speech with speakers of English and Spanish. *Cognition*. 2008; 109:274–280. [PubMed: 18842259]
- Brown-Schmidt S, Tanenhaus MK. Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*. 2006; 54:592–609.
- Cooper, WE.; Ross, JR. World order. In: Grossman, RE.; San, LJ.; Vance, TJ., editors. *Papers from the parasession on functionalism*. Chicago: Chicago Linguistic Society; 1975. p. 63-111.
- Davidson, DJ.; Bock, K.; Irwin, DE. Tick talk. In: Alterman, R.; Kirsh, D., editors. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. Boston: Cognitive Science Society; 2003. p. 294-299.
- de Smedt, K. Computational models of incremental grammatical encoding. In: Dijkstra, A.; de Smedt, K., editors. *Computational psycholinguistics: AI and connectionist models of human language processing*. London: Taylor & Francis; 1996. p. 279-307.
- Ertel, S. Where do the subjects of sentences come from?. In: Rosenberg, S., editor. *Sentence production: Developments in research and theory*. Hillsdale, NJ: Erlbaum; 1977. p. 141-167.
- Ferreira, F.; Engelhardt, PE. Syntax and production. In: Traxler, M.; Gernsbacher, MA., editors. *Handbook of Psycholinguistics*. New York: Academic Press; 2006. p. 61-92.
- Ferreira VS, Dell GS. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*. 2000; 40:296–340. [PubMed: 10888342]
- Flores d'Arcais, GB. Some perceptual determinants of sentence construction. In: Flores d'Arcais, GB., editor. *Studies in perception: Festschrift for Fabio Metelli*. Milan, Italy: Martello-Giunti; 1975. p. 344-373.
- Gleitman LR, January D, Nappa R, Trueswell JC. On the *give* and *take* between event apprehension and utterance formulation. *Journal of Memory and Language*. 2007; 57:544–569. [PubMed: 18978929]
- Griffin ZM, Bock K. What the eyes say about speaking. *Psychological Science*. 2000; 11:274–279. [PubMed: 11273384]
- Griffin, ZM. Why look? Reasons for eye movements related to language production. In: Ferreira, F.; Henderson, J., editors. *The integration of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press; 2004. p. 213-247.
- Henderson, JM.; Ferreira, F. Scene perception for psycholinguists. In: Ferreira, F.; Henderson, J., editors. *The integration of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press; 2004. p. 1-58.
- Irwin, DE. Fixation location and fixation duration as indices of cognitive processing. In: Ferreira, F.; Henderson, J., editors. *The integration of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press; 2004. p. 105-133.

- Kelly MH, Bock JK, Keil FC. Prototypicality in a linguistic context: Effects on sentence structure. *Journal of Memory and Language*. 1986; 25:59–74.
- Kempen G, Hoenkamp E. An incremental procedural grammar for sentence formulation. *Cognitive Science*. 1987; 11:201–258.
- Kuchinsky, SE.; Bock, K.; Irwin, DE. Shaping the link between seeing and saying. Poster presented at the Third International Workshop on Language Production; Chicago, IL. 2006 August.
- Lashley, KS. The problem of serial order in behavior. In: Jefferies, LA., editor. *Cerebral mechanisms in behavior*. New York: Wiley; 1951. p. 112-136.
- Levelt WJM. The speaker's linearization problem. *Philosophical Transactions of the Royal Society of London*. 1981; B295:305–315.
- Lounsbury, FG. Transitional probability, linguistic structure, and systems of habit-family hierarchies. In: Osgood, CE.; Sebeok, TA., editors. *Psycholinguistics: A survey of theory and research problems*. Bloomington: Indiana University Press; 1965. p. 93-101.
- MacWhinney B. Starting points. *Language*. 1977; 53:152–168.
- Madhavan, P.; Gonzalez, C. Effects of sensitivity, criterion shifts, and subjective confidence on the development of automaticity in airline luggage screening. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*; San Francisco. 2006 October. p. 334-338.
- Meyer, AS. The use of eye tracking in studies of sentence generation. In: Ferreira, F.; Henderson, J., editors. *The integration of language, vision, and action: Eye movements and the visual world*. New York: Psychology Press; 2004. p. 191-211.
- Onishi KH, Murphy GL, Bock K. Prototypicality in sentence production. *Cognitive Psychology*. 2008; 56:103–141. [PubMed: 17631877]
- Osgood, CE. Where do sentences come from?. In: Steinberg, DD.; Jakobovits, LA., editors. *Semantics: An interdisciplinary reader in philosophy, linguistics, and psychology*. Cambridge, England: Cambridge University Press; 1971. p. 497-529.
- Osgood, CE.; Bock, JK. Salience and sentencing: Some production principles. In: Rosenberg, S., editor. *Sentence production: Developments in research and theory*. Hillsdale, NJ: Erlbaum; 1977. p. 89-140.
- Paul, H. The sentence as the expression of the combination of several ideas. In: Blumenthal, AL., translator. *Language and psychology: Historical aspects of psycholinguistics*. New York: Wiley; 1970. p. 20-31. (Original work published 1886)
- Potter MC. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*. 1976; 2:509–522. [PubMed: 1003124]
- Schneider W, Shiffrin RM. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*. 1977; 84:1–66.
- Shiffrin RM, Schneider W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*. 1977; 84:127–190.
- Slobin, DI. From “thought and language” to “thinking for speaking.”. In: Gumperz, J.; Levinson, SC., editors. *Rethinking linguistic relativity*. Cambridge, UK: Cambridge University Press; 1996. p. 70-96.
- Tomlin, R. Focal attention, voice, and word order: An experimental, cross-linguistic study. In: Noonan, M.; Downing, P., editors. *Word order in discourse*. Amsterdam: John Benjamins; 1995. p. 521-558.
- Tomlin, RS. Mapping conceptual representations into linguistic representations: The role of attention in grammar. In: Nuyts, J.; Pederson, E., editors. *Language and conceptualization*. Cambridge, UK: Cambridge University Press; 1997. p. 162-189.
- van Nieuwenhuis KY, Dietrich R. Task sensitivity of animacy effects: Evidence from German picture description. *Linguistics*. 2003; 41:825–849.
- Yarbus, AL. *Eye movements and vision*. New York: Plenum Press; 1967.

Figure 1.
An example of an analog clock stimulus. Under regular instructions it is read as *eleven-forty* while under inverted instructions it is read as *eight fifty-five*.

Figure 2.

Eye movement patterns. First gaze onsets (left edge of bars) and durations (length of bars) to little and big hands. Time represented as a percentage of time between display onset and voice onset. (a) Regular time-telling. (b) Inverted time-telling. (c) Overlay of regular (solid bars) and inverted (striped bars) time-telling.

Figure 3. Eye-voice spans for first and second terms of time expressions. Error bars represent the standard error of the mean.

Table 1
Mean voice onset, gaze onset, and gaze duration times (msec) for hour and minute referents in both regular and inverted time-telling

	<u>Time-Telling Instructions</u>			
	Regular		Inverted	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Voice onset time	1626	75	1887	133
Hour referent gaze onset time	359	19	364	19
Hour referent gaze duration time	720	43	690	43
Minute referent gaze onset time	746	33	851	48
Minute referent gaze duration time	325	10	375	38