# Scan patterns on visual scenes predict sentence

# Scan Patterns predict Sentence Production in the Cross-modal Processing of Visual Scenes

Moreno I. Coco and Frank Keller
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
Phone: +44 131 650 4407, Fax: +44 131 650 4587
keller@inf.ed.ac.uk, mcoco@inf.ed.ac.uk

Abstract

Most everyday tasks involve multiple modalities, which raises the question of how the processing of these modalities is coordinated by the cognitive system. In this paper, we focus on the coordination of visual attention and linguistic processing during speaking. Previous research has shown that objects in a visual scene are fixated before they are mentioned, leading us to hypothesize that the scan pattern of a participant can be used to predict what they will say. We test this hypothesis using a data set of cued scene descriptions of photo-realistic scenes. We demonstrate that similar scan patterns are correlated with similar sentences, within and between visual scenes; and that this correlation holds for three phases of the language production process (target identification, sentence planning, and speaking). We also present a simple algorithm that uses scan patterns to accurately predict associated sentences by utilizing similarity-based retrieval.

*Keywords:* Scan patterns; eye-movements; language production; scene understanding; cross-model processing; similarity measures

## Introduction

Most everyday tasks humans perform involve multiple modalities. In order to successfully complete such tasks, information processing in all relevant modalities needs to be coordinated by the cognitive system. For example, the actions involved in making tea (finding the kettle, transporting it to the sink, locating and turning on the tap, etc.) involve a close interplay of visual attention and motor control, and eye-tracking studies indicate that fixations on task-relevant objects are coordinated with the appropriate motor actions, and typically precede them by around 600 ms (Land,

The support of the European Research Council under award number 203427 "Synchronous Linguistic and Visual Processing" is gratefully acknowledged.

Mennie, & Rusted, 1999). Furthermore, successful task completion requires that motor actions occur in a certain order, hence scan patterns (i.e., the temporal sequences of fixations across spatial locations) show a high degree of consistency across participants during everyday tasks such as walking, driving, playing ball games, or preparing food (see Land, 2006, for a review).

Language production is also an everyday task, and one that often happens in a visual context (giving directions on a map, explaining the function of a device, describing a picture). Cross-modal coordination is required in order to successfully construct linguistic descriptions in a visual context: objects need to be located in the scene, their visual features extracted, and relevant context information retrieved (e.g., for disambiguation: *the book on the floor* vs. *the book on the desk*).

Studies on situated language production show that visual objects tend to be fixated around 900 ms before their linguistic mention (Griffin & Bock, 2000), and that the production process (e.g., the selection of referents and the associated syntactic environment) is guided by visual information processing: a brief flash on a target location, prior to scene onset, favors its linguistic encoding (Gleitman, David January, Rebecca Nappa, & Trueswell, 2007).

Results such as these indicate that cross-modal coordination in language production occurs based on shared referential information. Visual referents, such as objects (Griffin & Bock, 2000) or locations (Gleitman et al., 2007), trigger the production of the corresponding linguistic referents. Furthermore, we can hypothesize that the ordering of referents is exploited for cross-model coordination: both the visual processing stream (scan patterns performed on a scene) and the linguistic one (sentences uttered) are sequentially ordered. The order of referents is in fact a crucial determinant of the meaning of an utterance (*dog bites man* vs. *man bites dog*).

If referents, and the order they occur in, are at the heart of the coordination between visual processing and language production, then it should be possible to exploit this relationship directly. In particular, it should be possible to retrieve a sentence a participant will say based on the associated scan pattern information.

The aim of the present paper is to test this hypothesis. More specifically, we will first show that scan pattern similarity and sentence similarity are correlated in a sentence production task, and subsequently demonstrate that it is possible to utilize cross-modal similarity to predict the sentence a participant produced based on the associated scan pattern. In the context of this paper, prediction is operationalized as a retrieval. Thus, by "prediction" we mean that the correct sentences is selected from a pool of available utterances based on the associated scan pattern. Our sentence production task requires participants to generate a verbal description of a photo-realistic scene, after being prompted with a cue word that refers to one of the objects in the scene. This task has similarities with visual search, a well-studied task in the visual cognition literature. We can assume that the cue for the scene description triggers a search for the corresponding object in the scene. Once the target has been found, further search processes are initiated to identify other objects that need to be mentioned to produce a well-formed sentence. These search processes are driven by contextual expectations arising from the sentence the participant is planning to utter, and from the properties of the objects he or she has already identified.

Such context effects are well-established in visual search experiments. When participants inspect naturalistic scenes in order to locate an object, their attention is guided by expectations about likely target positions (Findlay & Gilchrist, 2001; Neider & Zelinsky, 2006; Henderson, 2007). In indoor scenes, for example, mugs are likely to be located on desks or worktops, while paintings are likely to be found on walls. Such contextual expectations are computed at the first fixation on a scene (Potter, 1976; Vo & Henderson, 2010) and constrain subsequent fixations to regions that

are contextually relevant (Torralba, Oliva, Castelhano, & Henderson, 2006). We expect context to play a similar role in a cross-modal task such as scene description, in particular as there is evidence that visual search can avail itself of linguistically specified context information. For example, a verbal description of the search target conveys an advantage over a pictorial specification (Yang & Zelinsky, 2009), with more detailed descriptions conveying a larger advantage (Schmidt & Zelinsky, 2009). Furthermore, Griffin and Bock's ((2000)) study finds that eye-movements during verbal event description and (non-verbal) event comprehension are similar, but differ from those during free viewing. These results confirm the role of contextual and task factors in visual processing, and indicate that these factors are also an important determinant of eye-movements during linguistic processing.

Contextual factors affect not only the location and duration of fixations, but also their sequence (the scan pattern). In a scene memorization experiment, Foulsham and Underwood (2008) exposed participants to the same scene twice (once for encoding and once for recognition) and found that the resulting scan patterns were more similar than if participants were exposed to two different scenes. Scan pattern similarity is increased further if participants are engaged in the same task (encoding/recognition vs. imagery), as Humphrey and Underwood (2008) demonstrated. These results for single-modality processing are confirmed by studies using cross-modal tasks, such as making tea (Land et al., 1999) or preparing a sandwich (Hayhoe, 2000). In these tasks, scan patterns are observed to be highly similar across participants, presumably because eye-movements have to be coordinated with sequences of motor actions, which are largely predetermined by the nature of the task. We expect this finding to carry over to the cross-modal task under investigation in the present paper, scene description, which typically requires a close coordination of eye-movements and linguistic processing.

Previous research investigating this coordination has mostly employed the Visual World Paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), in which participants listen to a speech stimulus while viewing a visual context (typically an array of objects). Research using this paradigm has demonstrated a clear link between the processing of certain linguistic constructions and attention to contextually relevant visual information (e.g., Knoeferle & Crocker, 2006). However, most visual world studies have focused on specific psycholinguistic phenomena (e.g., attachment ambiguity) rather than on general mechanisms underlying the coordination between scene understanding and sentence processing.

The aim of the present paper is to unify previous evidence about the interaction of visual and linguistic processing under the theoretical framework of cross-modal coordination. We demonstrate the validity of this framework by testing the hypothesis that scan pattern information is predictive of sentence production. In particular, we assume cross-modal coordination of visual and linguistic processing to be primarily based on referential overlap, i.e., the objects looked at are associated with the words mentioned. Moreover, we expect cross-modal coordination to unfold sequentially across the two streams, i.e., objects are looked at before being mentioned.

Note that the existing visual world literature does not provide direct evidence to this hypothesis, as the cross-modal similarity of visual and linguistic information has not being quantified before. Moreover, in contrast with previous research, we are able to apply the theoretical principle of cross-modal coordination and implement a retrieval algorithm that is able to correctly identify a sentence on the basis of the associated scan pattern information, among all other possible sentences.

Beside advancing the understanding of situated language production theoretically, our results introduce a novel and more integrated methodological approach to the analysis of visual and linguis-

tic information. In fact, most of the published visual world experiments use objects arrays (usually, 3-5 objects) or simple scenes (line drawings or clip art). Even if scenes are used, these tend to be highly artificial and their referential information is often contextually unrelated (e.g., scenes involving a WIZARD and a BALLERINA in Knoeferle & Crocker, 2006). Instead, we use photo-realistic scenes and fully explore their complexity by studying scan patterns rather than looks to specific regions of interest. We are aware of one other recent study (Andersson, Ferreira, & Henderson, 2011) using photo-realistic scenes, but this work does not go beyond the analysis of specific target regions, therefore merely confirms the results of existing studies with image arrays containing a large number of objects.

Also the range of linguistic structures investigated in VWP literature is comparatively small (e.g., Griffin & Bock, 2000, only looked at actives and passives), while our aim is to investigate a varied range of productions. Similarly to the scan pattern information, we will explore the sentence information as a whole, rather than focusing on specific target words.

In a nutshell, in order to obtain a maximally general test of the hypothesis that scan patterns predict sentence productions based on the principle of cross-modal coordination, this paper employs data from a cued sentence production experiment with a diverse set of photo-realistic scenes containing a large number of objects, where participants were free to produce any scene description they liked.

## Method

The aim of this experiment was to test the claim that scan patterns on visual scenes are predictive of sentence productions. More specifically, we investigated whether there is an association between scan pattern similarity and sentence similarity, using both correlation analysis and mixed effects modeling.[1] Moreover, in order to strengthen the theoretical validity of our results and demonstrate a concrete application of our work, we develop an algorithm able to accurately retrieve a sentence based on the associated scan pattern.

### Data Collection and Pre-processing

The data used in this study was that of Coco and Keller's ((2010)) language production experiment. In this experiment, 24 participants were eye-tracked while describing photo-realistic indoor and outdoor scenes. The experimental materials consisted of 24 different scenes, all of which depicted indoor environments, drawn from six different scenarios (e.g., bedroom, entrance). The experiment also included 48 fillers.

For each scene, participants were prompted with a cue word which referred to a visual object in the scene, and were told to use the cue word in their description. The cue was presented in written form at the beginning of the trial for 750 ms before the onset of the scene. The cue words were either animate or inanimate (e.g., *man* or *suitcase*) and were ambiguous with respect to the scene, i.e., they could refer to more than one depicted object (see Fig. 1 for an example trial).

Participants' eye-movements were recorded using an SR Research Eyelink II eye-tracker with sampling rate of 500 Hz on a 21" screen (1024 $\times$ 768 pixel resolution), while the speech of the participants was recorded with a lapel microphone.

[1] We only report the results for one similarity measure here; corroborating results involving a larger range of similarity measures are presented in Appendix A.


*Figure 1*. Example of scene and cue words used as stimuli for the description task. Each scene was manually segmented into polygons, drawn around objects using the LabelMe toolbox (Russell, Torralba, Murphy, & Freeman, 2008). Each polygon was annotated with the corresponding linguistic label.

The resulting data set contains 576 sentences produced for the 24 scenes. The sentences were manually transcribed and paired with the scan patterns that participants followed while generating them. Two pairs were discarded because the sentences were missing. Each scene was fully annotated using the LabelMe toolbox (Russell et al., 2008) by drawing bounding polygons around the objects in the scene and labeling them with words (see Fig. 1). These polygons were used to map the fixation coordinates onto the corresponding labels. Objects can be embedded into other objects (e.g., the head is part of the body); in this case, the smallest object that contained the fixation was used. The mean number of objects per image was 28.65 (SD = 11.30).

A scan pattern is represented as a sequence of object labels encoding the objects that were looked at in the temporal order of fixation. There is substantial variation in the data set both in terms of the complexity of the sentences that were produced (e.g., *one man waits for another man to fill out the registration form for a hotel* vs. *the man is checking in* for Fig. 1) and in terms of the length of the scan patterns observed prior to production (min = 800 ms; max = 10205 ms) and during production (min = 2052 ms; max = 18361 ms). To account for this variability, we use a measure of sentence and scan pattern similarity that is insensitive to length.

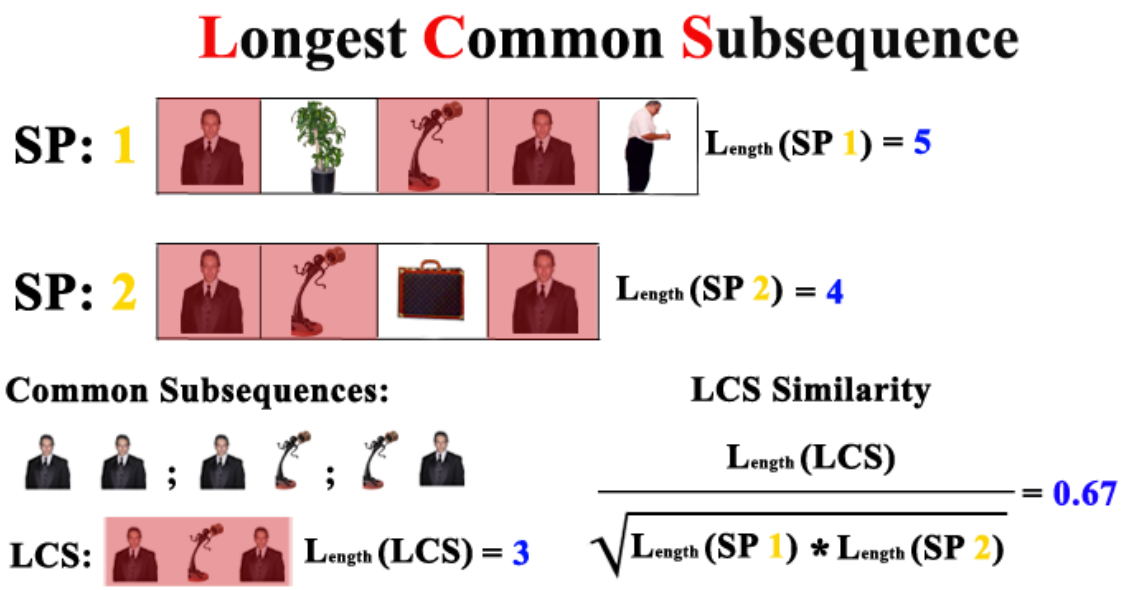



*Figure 2*. Longest Common Subsequence is a measure of similarity based on ordered subsequences. For two sequences, it explores all common subsequences and returns the longest one. Here, SP1 and SP2 share several common subsequences of length 2 (e.g., man–man). The longest common subsequence is man–statue–man, i.e., of length 3.

### Similarity Measure

Before we can quantify the association between scan patterns and sentence productions, we need to define a similarity measure. We use Longest Common Subsequence (LCS, Gusfield, 1997), which can be used to quantify similarity both for scan patterns and word sequences. (Please refer to Appendix A for additional analyses using other measures.)

Both scan patterns (sequences of fixated objects) and sentences (sequences of words) are sequential data. Finding similarity between sequences is a well-known problem in bioinformatics (Durbin, Eddy, Krogh, & Mitchison, 2003), where genetic codes have to be compared to unravel underlying similarities. A guiding principle used to capture these similarities is *alignment*. The more elements can be aligned across the two sequences, the more similar they are. Two key issues have to be overcome in order to compute alignment: sequences differ in length, and the elements in the sequences, even if identical, can be positioned differently. LCS addresses these issues.[2]

LCS finds the longest subsequence of two sequences. The LCS algorithm searches the space of all combinations of ordered subsequences, looking for the alignment which maximizes the number of common elements. The algorithm follows a dynamic programming approach, where the final solution (the longest alignment) is iteratively built based on solutions for subproblems (alignments of shorter subsequences). Once the longest subsequence is found, the similarity score is calculated as the ratio between the length of longest common subsequence and the geometric mean of the lengths of the two sequences. The resulting values range from 1 for most similar to 0 for least sim-

[2]It has recently been proposed to apply a standard bioinformatics algorithm (the Needleman-Wunsch algorithm) to eye-movements data (ScanMatch, Cristino, Mathot, Theeuwes, & Gilchrist, 2010). For our data, LCS and ScanMatch are highly correlated ($\rho = 0.98; p < 0.001$), so we only report the results obtained using the simpler LCS method.

ilar. Fig. 2 gives an example: SP1 and SP2 share several common subsequences, e.g., man–man and man–statue, with a length of 2. The algorithm explores all possible combinations and finds the longest common subsequence, in this case man–statue–man. Often, LCS finds more than one solution, i.e., there are several common subsequences of maximal length; this has no effect on the similarity score.

**Data Analysis**

LCS sequence analysis can be applied on both scan pattern and sentence data. We applied LCS on sentences (LCS.L) with low frequency words removed, and on scan patterns without time information (LCS.V). (Results for a similarity measure that includes time information are presented in Appendix A.)

To analyze the correspondence between sentences and scan patterns, we divide the data into three regions: Planning, Encoding, and Production. The Planning region starts with the onset of the image and lasts until the cued object is fixated. This phase essentially captures the visual search taking place to find the cue. The Encoding region starts with the first fixation on the cued object and runs until speech onset; this phase captures the information retrieval behavior that happens in support of sentence encoding. Finally, the Production region runs from the beginning to the end of the speech. In this region, linguistic and visual information processing happen concurrently.

For each region of analysis, LCS is computed pairwise, i.e., every trial (sentence and scan pattern) is paired with every other trial. As an example, if participant 1 generates the sentence *the man is reading a newspaper*, we calculated the similarity of this sentence with any other sentence in the dataset (e.g., *the man is washing an apple* from participant 2). The similarity score was calculated in the same way for the associated scan patterns. This resulted in a total of 382,973 pairs over the three different regions of analysis (Planning = 123,256, Encoding = 95,266, and Production = 164,451). The difference in sample size for the regions is due to temporal overlap. For instance, some participants start speaking before having fixated at the target object (i.e., no Encoding), or fixate the target at the onset of scene (i.e., no Planning). For Production we have the highest number of pairs, as there is no overlap with other regions, by definition. Note that sentence similarity for the sentence is constant across the three regions of analysis; only the scan pattern information changes. This means that the same LCS.L similarity scores are paired with three different, region specific, LCS.V scores.

We first explore the pattern of correlations across regions. We then provide a more detailed analysis of these correlations by using linear mixed effect modeling (LME, Baayen, Davidson, & Bates, 2008), an approach which allows us to take random factors (such as between-participant and between-trial variation) into account.

We investigate the data at two levels: (1) globally, i.e., by performing comparisons between all pairs of trials in the full data set, and (2) locally, i.e., by comparing only the trials that pertain to the same scene (24 in total). These two levels of analysis make it possible to test whether the coordination between sentences and scan patterns is scene specific. We also report a baseline correlation (Foulsham & Underwood, 2008) that is obtained by pairing sentences and scan patterns randomly (rather than pairing the scan patterns with the sentences they belong to). We quantify the strength of the correspondence between similarity measures by computing Spearman's $\rho$ for all pairs of measures.

The distinction between global and local similarity has implications for the nature of cross-modal coordination. A correlation found globally (across all scenes) would imply that scan patterns

are partially independent of the precise spatial configuration of the scene (the position of the objects, etc.), as this varies across scenes. The correlation would instead be driven by the referential structure that is shared across scenes, i.e., by the identity of the objects and the order in which they are mentioned or fixated. A correlation at the local level would be consistent with well-known scene-based effects, both bottom-up and top-down, which guide visual attention (Itti & Koch, 2000; Foulsham & Underwood, 2008).

An important aspect of cross-modal coordination is the role of ordering. We want to test the hypothesis that cross-modal coordination emerges as the product of referential overlap and sequential ordering. In order to test whether both of these components play a role in cross-modal coordination, we compare LCS, which is a measure sensitive to ordering, with a Bag of Words (BoW) similarity measure, both for sentences and scan patterns. The BoW measure counts the number of common elements (objects or words) in two sequences, over the total number of elements. This means BoW gives us a measure of referential overlap, without taking ordering into account. We assess the significance of difference in correlation coefficients between LCS and BoW after applying a Fisher z-transformation of the coefficients. (A more detailed analysis of this comparison, focusing on the role of sample size, can be found in Appendix A.)

In the linear mixed effects analysis, we used scan pattern similarity as the dependent variable and sentence similarity as the predictor. Region of analysis (Planning, Encoding, and Production) and Cue (Animate, Inanimate, Mixed) were included as other factors. Regions of analysis is contrast coded, with Encoding as the reference level, while the reference level for Cue is Inanimate. Note, Cue has three levels because of pairwise comparison, i.e., we compare also Animate with Inanimate cases (Mixed).

The random factors were participants and trials, random slopes under sentence similarity and region were also included. A forward selection procedure was used to obtain the minimal model based on these random and fixed factors (see Appendix B for details on model selection).

## Results and Discussion

Fig. 3 plots the linguistic similarity measure LCS.L against the scan pattern similarity measure LCS.V, computed globally, i.e., across all scenes. We observe a clear association between sentence and scan pattern similarity: when LCS.L values increase, LCS.V values also increase. This effect is consistent and statistically significant across all three regions of analysis (Planning: $\rho = 0.48$; Encoding: $\rho = 0.47$, Production: $\rho = 0.38$; $p < 0.001$ in all cases), but does not occur in the random baselines ($\rho \approx 0.002$; $p > 0.1$ in all cases).

For BoW similarity, we also find a clear association between sentence and scan pattern similarity across all three regions of analysis (Planning: $\rho = 0.30$; Encoding: $\rho = 0.43$, Production: $\rho = 0.34$; $p < 0.001$ in all cases). However, the correlation coefficients for BoW are smaller than ones obtained for LCS across all three phases; the difference is particularly large for Planning ($z = -52.99$; $p < 0.001$), but also significant for Encoding ($z = -8.25$; $p < 0.001$ ) and Production ($z = -9.92$; $p < 0.001$); refer to the Appendix A for a more detailed full analysis. We report LME analysis using LCS only, as it seems to better capture the cross-modal coordination, compared to BoW.

Fig. 4 plots local similarity values, i.e., LCS.V and LCS.L values computed separately for each scene. The trend observed at the global level is confirmed across all regions, though there is substantial variation in the degree of association between scan pattern and linguistic similarity from scene to scene. Tab 1 gives the minimum and maximum values of the correlation coefficients within

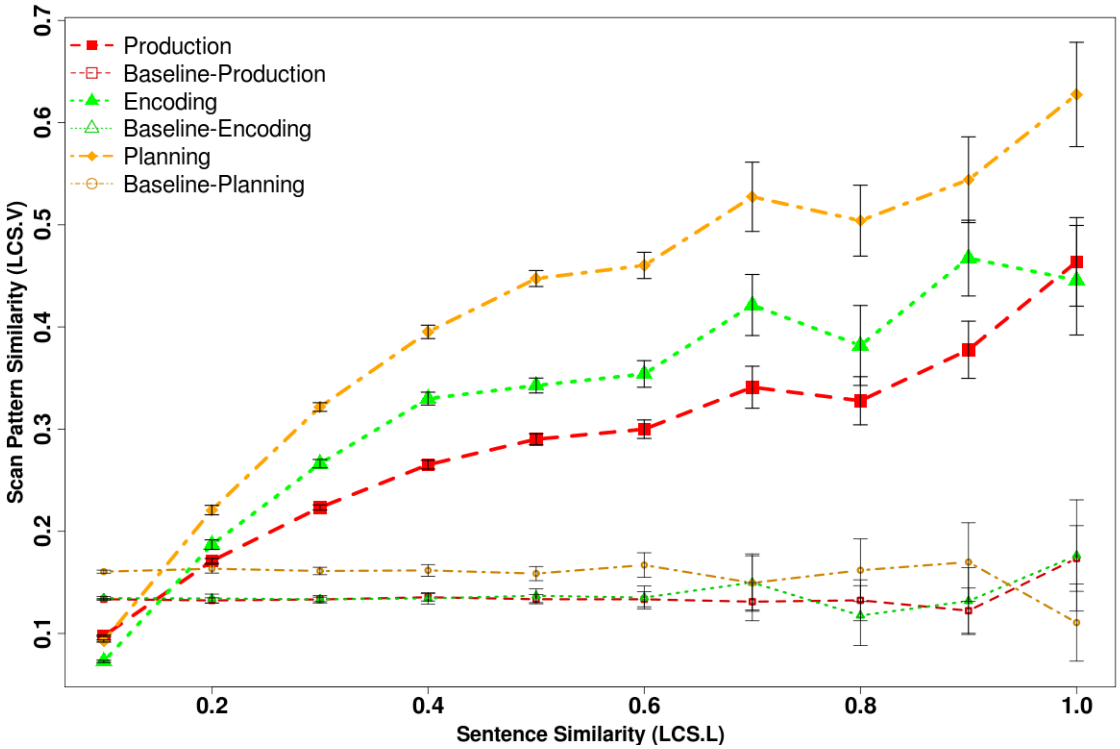



*Figure 3*. Scan pattern similarity (LCS.V) as a function of linguistic similarity (LCS.L) across all 24 scenes. The data has been binned on the x-axis; the whiskers represent 95% confidence intervals. The baselines were obtained by randomly pairing a sentence and a scan pattern from the respective phase.

scenes. As expected from Fig. 4, correlation coefficients vary across scenes for all pairs of measures, which indicates that scene context modulates the coordination between scan patterns and linguistic productions.

Turning now to the linear mixed effects analysis, Fig. 5 shows a plot of LME predicted values calculated globally (for the LME coefficients refer to Tab 2, see also the discussion below). The model closely follows the empirical patterns in Fig. 3 with the scatter indicating a positive association: scan pattern similarity increases with linguistic similarity. Note that the majority of data has a low overall similarity (many observations lie between 0 and 0.4). However, there is another cloud of data points, similarly distributed, with higher LCS.V values overall. On closer inspection, it becomes clear that data points with higher similarity are based on within-scene comparisons, while those with lower similarity are based on between-scene comparisons.

To investigate this observation further, Fig. 6 plots the density of cross-modal similarity aggregated separately within scenes, i.e., only for trials from the same scene, and between scenes, i.e., only for trials from different scenes. Fig. 6 shows that the resulting similarity scores are normally distributed and higher within scenes than between scenes (where the distribution is also more skewed).

Tab 2 lists the coefficients of the mixed effects models; these are consistent with the descrip-

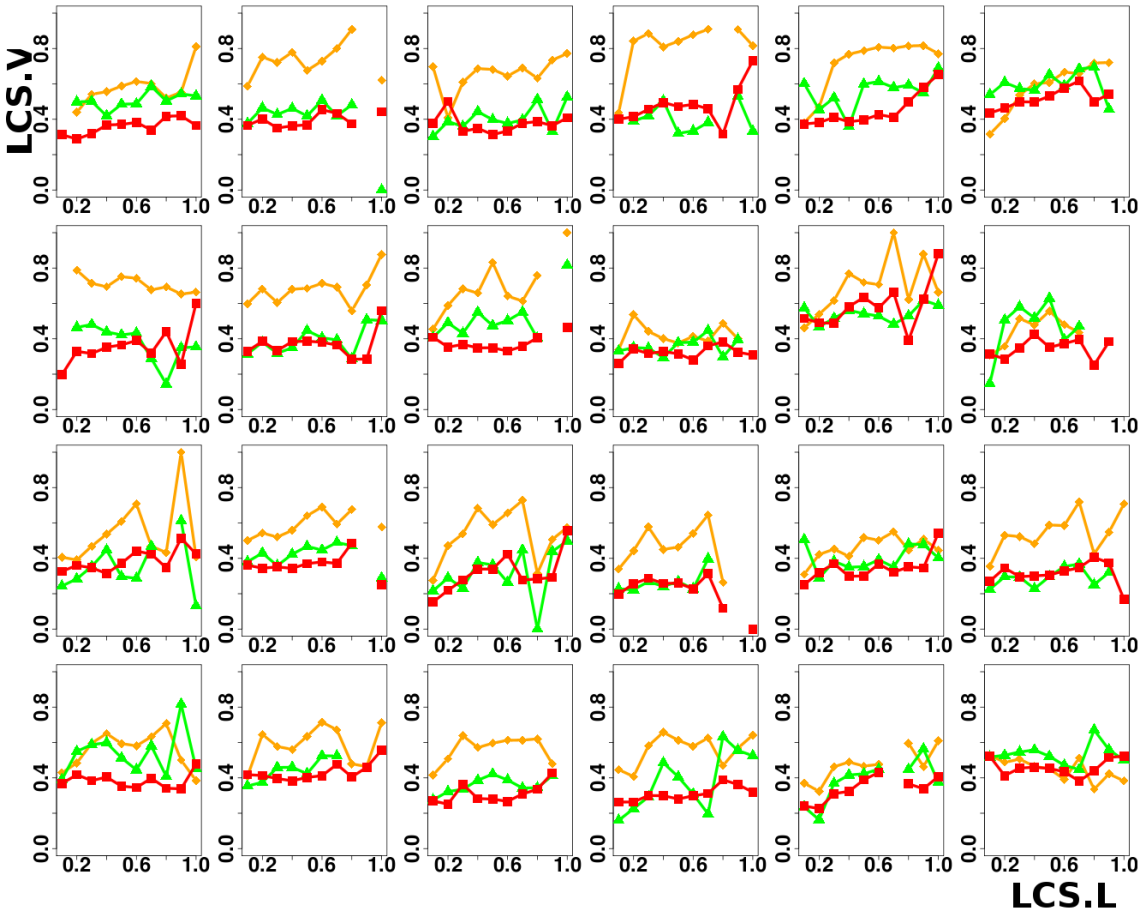



*Figure 4*. Scan pattern similarity (LCS.V) as a function of linguistic similarity (LCS.L), separately for each of the 24 scenes. Line color and the point type represent the different regions of analysis: Planning (orange, circle), Encoding (green, triangle), Production (red, square)

Table 1

*Minimum and maximum correlations (Spearman* ρ*) between scan pattern similarity (LCS.V) and linguistic similarity (LCS.L), across regions of analysis (Planing, Encoding, Production); correlations were computed for each scene separately. All correlations are significant at* $p < 0.05$.

| | Planning | Encoding | Production |
|---|---|---|---|
| Min | 0.25 | 0.14 | 0.04 |
| Max | 0.63 | 0.77 | 0.51 |

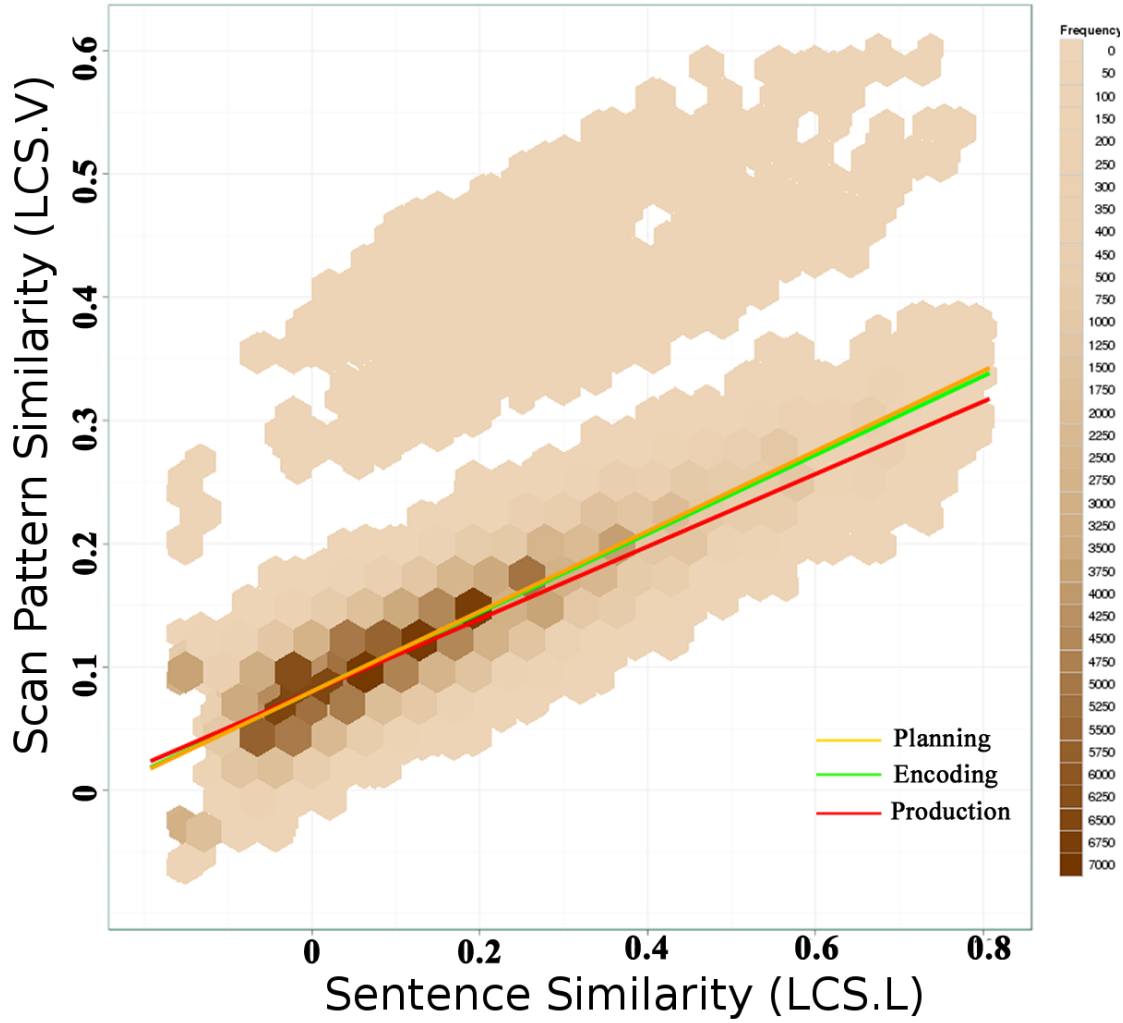



*Figure 5*. Hexagonal plot of predicted values of the linear mixed effects model: linguistic similarity predicted by scan pattern similarity. The plot shows the observed data binned into hexagons. The color of the hexagon reflects the frequency of the observations within it (darker for more observations). The solid lines represent the grand mean intercept and slope: Planning (orange), Encoding (green), Production (red).

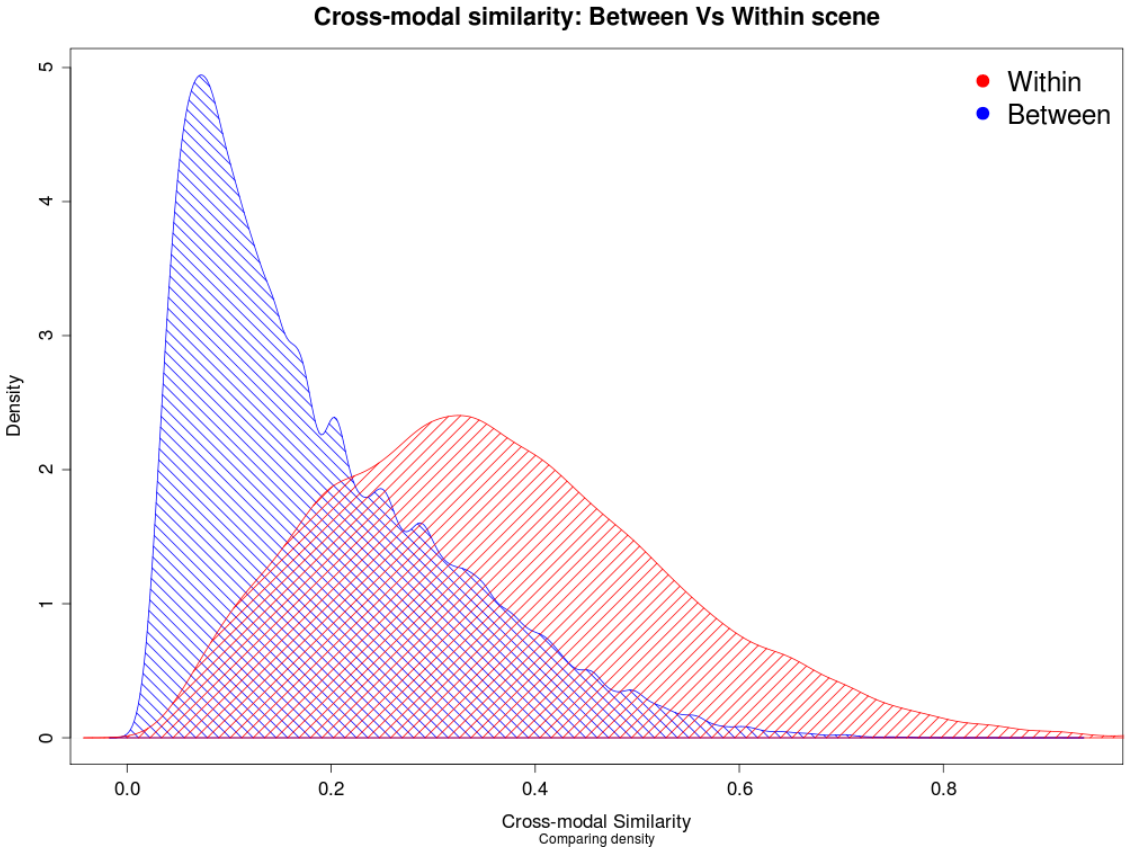



*Figure 6*. Density plot for cross-modal similarity. Cross-modal similarity is computed by summing the similarity scores obtained separately for the linguistic and scan pattern measure and normalizing them to a range between 0 and 1. Trials with a cross-modal similarity of 0 were removed. Red line: cross-modal similarity within the same scene, blue line: between different scenes.

tive results. We find a significant main effect of sentence similarity (LCS.L).[3] The coefficient is positive, which confirms our basic finding, i.e., that more similar sentence production lead to more similar scan patterns. There are also significant main effects of Regions: $\text{Region}_{\text{PlanVsEnc}}$ has a positive coefficient, which means that overall similarity is higher in the Planning region compared to the Encoding region, while $\text{Region}_{\text{ProdVsEnc}}$ has a negative coefficient: the similarity is lower in Production compared to Encoding. We also find significant main effects of Cue, with $\text{Cue}_{\text{AniVsIna}}$ having a negative coefficient, which means that Animate targets triggers lower similarity than Inanimate one, which shows higher similarity even compared to the mixed case ($\text{Cue}_{\text{MixVsIna}}$ has a positive coefficient).

Turning on the interactions, we find significant interactions of LCS.L with both Cue and Region. In particular, LCS.L positively interacts with Cue (coefficient LCS.L:$\text{Cue}_{\text{AniVsIna}}$), indicat-

[3] In correlation analysis the direction of causality cannot be inferred. When we re-do the analysis with scan pattern similarity as a predictor and sentence similarity and the dependent variable, we obtain the same result: sentence similarity increases when scan pattern similarity also increases.

ing that most similar sentences tend to be associated with Animate targets. Regarding Regions, it emerges that sentence similarity increases more sharply with the scan pattern similarity in the Planning region than in the Encoding region, while the opposite is found for Production (coefficient LCS.L:$\text{Region}_{\text{ProdVsEnc}}$). A significant interaction is found also for the factors Cue and Region, and the results are in line with the interactions observed with LCS.L. In particular, sentences associated with Animate targets are less similar in Production ($\text{Cue}_{\text{AnivsVsIna}}$:$\text{Region}_{\text{ProdVsEnc}}$), with Inanimate again more similar compared to the mixed case ($\text{Cue}_{\text{MixvsVsIna}}$:$\text{Region}_{\text{ProdVsEnc}}$). Animate targets instead lead to higher similarity in Planning compared to Encoding.

The LME analysis therefore confirms the trend visible in Fig. 3, and corroborates the more basic correlation analysis presented earlier. Beside providing evidence for our main hypothesis that scan patterns predict sentence production, our results shed light on how cross-modal similarity interacts with the different phases of the task, and with the animacy of the cue word.

In particular, the change of similarity observed across the different phases can be explained if we consider that scan pattern similarity is a function of the amount of visual processing involved. At the beginning of a trial, a search for the cued target object is launched, and the visual system allocates attention solely based on scene information, leading to high scan pattern similarity (Henderson, 2007; Malcolm & Henderson, 2010). During Encoding and Production, however, scan patterns are increasingly determined by linguistic processing, as the objects that are to be mentioned in the sentence are selected and sequenced. This means that general scene information becomes less relevant, reducing the overall amount of scan pattern similarity. Moreover, these patterns vary contingently with the type of associated sentence. Sentences containing an Animate referent have to be truly similar in order to be associated with similar scan patterns: an animate referent is usually situated within the broader contextual information (e.g., *the man is signing in*) while an inanimate referent has to be spatially located in relation to another ground object (e.g., *the suitcase is on the floor*). This means that small changes between two animate sentences (e.g., *the man is standing in the reception of a hotel*) can entail a large set of different visual referents in the associated scan patterns. In contrast, small changes between two inanimate sentences (e.g., *the suitcase is next to the man*, have a smaller effect on scan pattern similarity, as the set of objects spatially related to inanimate visual referents is relatively constrained.

## Application to Sentence Prediction

So far, we have demonstrated that the similarity of scan patterns is correlated with the similarity of the associated sentences, and we have established to what extent cross-modal similarity is mediated by the region of analysis and the animacy of the cue. In the following, we illustrate how the correlation between scan pattern similarity and sentence similarity can be used to predict which sentence a participant uttered based on which scan pattern they followed. Our aim is to provide a proof of concept, so we implement a simple algorithm that uses scan pattern information to retrieve sentences from a pool of candidates. In other words, we define prediction as sentence retrieval, rather than generating sentences from scratch based on scan pattern information.

The algorithm exploits the fact that similar scan patterns (e.g., as quantified by LCS.V) are associated with similar sentences (e.g., as measured by LCS.L). The algorithm is illustrated schematically in Fig. 7. We start by picking a scan pattern $V_S$ at random from the pool of all available scan patterns. The sentence associated with it is denoted by $L_S$. Then, based on the pairwise *visual* similarity scores, we extract the scan pattern $V_T$ that is most similar to $V_S$ from the pool (if several scan patterns share the maximal similarity score, we pick one at random). The sentence associated

Table 2

*Coefficients for the mixed effects model analysis. The dependent measure is scan pattern similarity (LCS.V); predictors are sentence similarity (LCS.L), Region (contrast coded with Embedding as the reference level) and Cue (contrast coded with Inanimate as the reference level). Note that Cue has three levels because of pairwise comparison, i.e., we compare also Animate with Inanimate (level Mixed). The abbreviations are: Plan: Planning; Enc: Encoding; Prod: Production; Ani: Animate; Ina: Inanimate; Mix: Mixed.*

| Predictor | Coefficient |
|---|---|
| (Intercept) | $0.188^{***}$ |
| LCS.L | $0.514^{***}$ |
| $\text{Region}_{\text{PlanVsEnc}}$ | $0.038^{***}$ |
| $\text{Region}_{\text{ProdVsEnc}}$ | $-0.018^{***}$ |
| $\text{Cue}_{\text{AniVsIna}}$ | $0.038^{***}$ |
| $\text{Cue}_{\text{MixVsIna}}$ | $0.005^{***}$ |
| $\text{LCS.L:Region}_{\text{PlanVsEnc}}$ | $0.326^{***}$ |
| $\text{LCS.L:Region}_{\text{ProdVsEnc}}$ | $-0.352^{***}$ |
| $\text{LCS.L:Cue}_{\text{AniVsIna}}$ | $0.093^{***}$ |
| $\text{Cue}_{\text{AniVsIna}}\text{:Region}_{\text{PlanVsEnc}}$ | $0.008^{***}$ |
| $\text{Cue}_{\text{AniVsIna}}\text{:Region}_{\text{ProdVsEnc}}$ | $-0.019^{***}$ |
| $\text{Cue}_{\text{MixVsIna}}\text{:Region}_{\text{ProdVsEnc}}$ | $0.014^{***}$ |

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

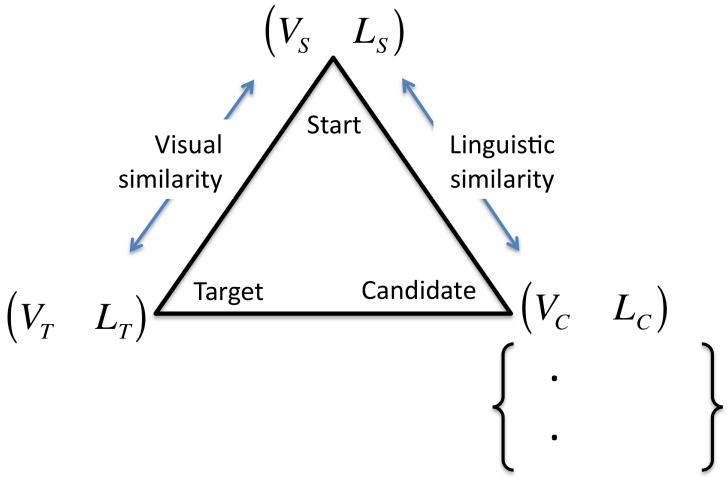



*Figure 7*. Schematic illustration of the sentence retrieval algorithm.

with $V_T$ is $L_T$. In light of the results of our correlation analyses, we expect the linguistic similarity between $L_S$ and $L_T$ to be high. Our goal is to retrieve $L_T$ from the pool of all available sentences. To measure how successful this is, we consider sets of sentence candidates $L_C$ of increasing size, where the target $L_T$ is always included in the set.

For each set size ($N = 2, \ldots, 572$), we compute the pairwise *linguistic* similarity scores between $L_S$ and $L_C$, and extract the best candidate sentence $L_C^*$, i.e., the one that maximizes this similarity. The underlying assumption is that if $L_C$ and $L_S$ are similar, by the transitive property of similarity, $L_C$ and $L_T$ must also be similar. We repeat this procedure for 50 iterations (this is necessary because of the random sampling involved), count how many times we have retrieved the correct sentence ($L_C^* = L_T$), and divide this number by the total number of iterations. This measure gives us the retrieval accuracy, which we compare against chance, i.e., the probability of finding the target at random in a set of size $N$. We apply the algorithm to the Planning, Encoding, and Production regions separately.

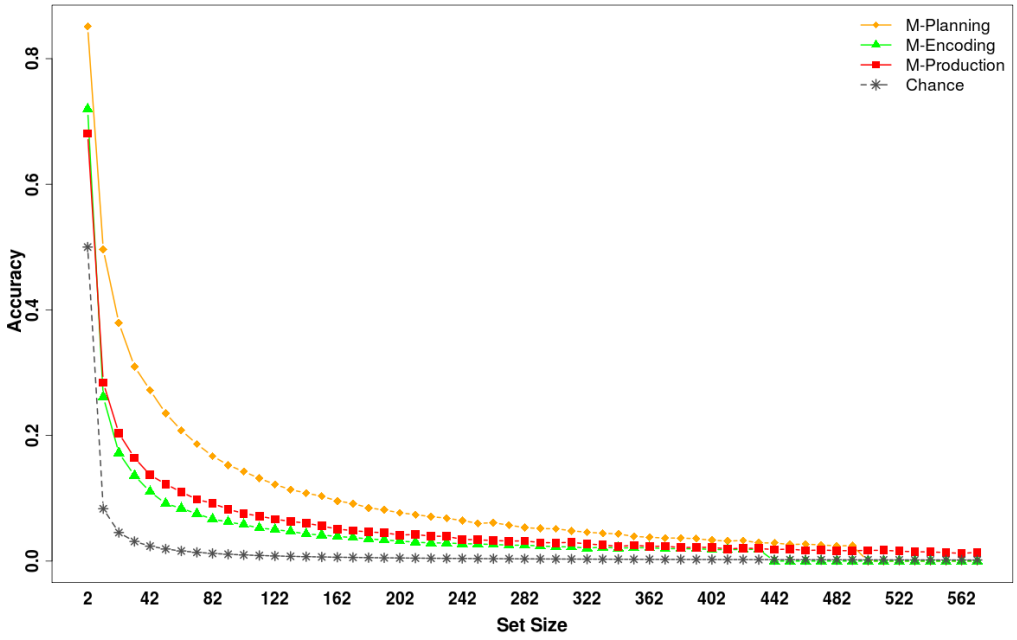
In Fig. 8(a), we plot the accuracy of our algorithm as a function of set size, for the three regions of analysis. The result shows that our algorithm performs significantly better than chance across all set sizes in retrieving the target sentence associated with scan pattern of interest. This is consistent across all regions of analysis, with Planning having the best accuracy, followed by Production and Encoding. This result is compatible with our previous correlation analysis, which showed a similar trend (see Fig. 3). As expected, the performance of the algorithm degrades with increasing set size, as the chance for encountering a sentence similar to the target increases.

Another aspect of the performance of the algorithm is the similarity between a predicted non-target sentence and the correct target sentence. This measures the performance in cases where the algorithm does not find the target. Again, sentence similarity varies with set size, see Fig. 8(b) (which excludes cases where the algorithm finds the target). We observe that mean linguistic similarity approaches a constant, which is higher for Encoding than for Planning and Production. This illustrates that the performance of the algorithm in retrieving similar sentences to the target does not deteriorate with set size, and the decreasing trend of Fig. 8(a) can be entirely attributed to increasing difficulty of identification the exact target sentences (rather than a similar one).

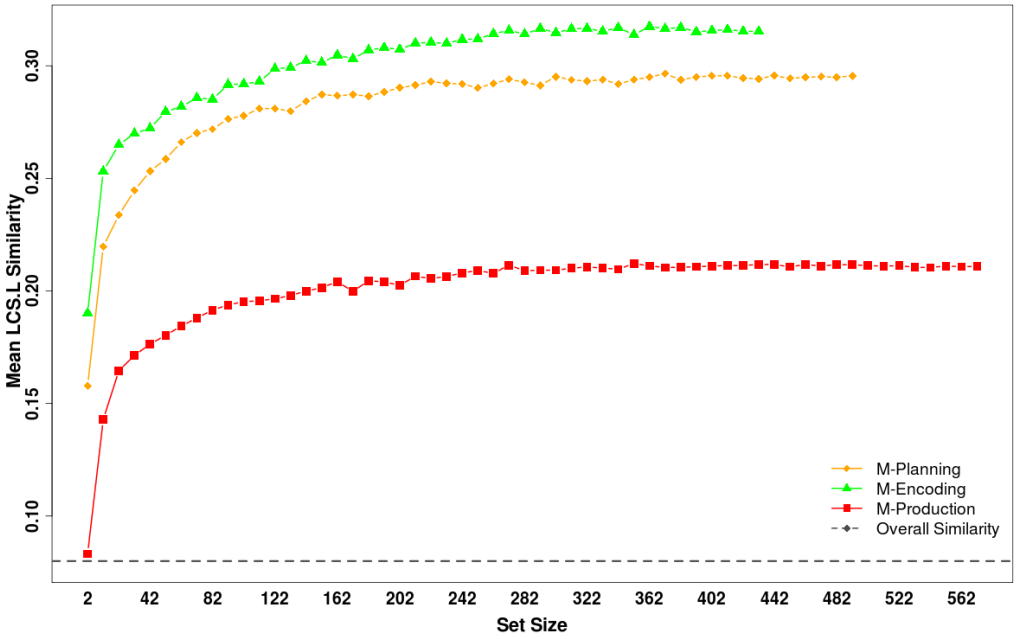
As mentioned in the beginning, this study is mainly meant as proof of concept. The performance of our algorithm could be optimized to achieve a higher accuracy, for example by using a combination of similarity measures as opposed to a single one.

## General Discussion

Cross-modal coordination is a key feature of the cognitive system, enabling it to successfully perform tasks that draw on information from multiple modalities. Previous research has shown, for instance, that efficient motor planning relies on the coordination between visual attention and action (Land et al., 1999; Hayhoe, 2000; Land, 2006). In this paper, we investigated the coordination between visual attention and linguistic processing during the description of naturalistic scenes, and found that scan pattern similarity is correlated with sentence similarity. This result was obtained both in a simple correlation analysis and in a mixed model which takes between-participant and between-trial variation into account. We divided the scene description task into three phases (Planning, Encoding, Production), and found that the association between scan patterns and sentences was consistent across these phases. Furthermore, the association is stable across a range of similarity measures (see Appendix A).




(a) Proportion of correctly retrieved sentences as a function of set size.




(b) Mean linguistic similarity between the target-sentence and the predicted sentences, as a function of set size.

*Figure 8*. Evaluation of retrieval algorithm in terms of accuracy and similarity.

On a theoretical level, our findings suggest that cross-modal coordination is primarily a *referential* phenomenon, i.e., that it is driven by the sequence of objects fixated in a scan pattern, and by the sequence of objects mentioned in the speech, respectively. This is corroborated by the fact that we found cross-modal coordination both within and between scenes; different scenes differ in visual features such as layout, object position, or size, but the referents (i.e., the identity of the objects) can overlap between scenes. Cross-modal coordination therefore goes beyond the coordination based on low-level visual features within a scene reported in previous work (Foulsham & Underwood, 2008).

Ordering is another important feature of cross-modal coordination, as the processing of objects (scan patterns) and words (sentence) unfolds sequentially. In fact, we found that similarity measures based only on referential overlap (BoW) are significantly less correlated than measures which are also partially sensitive to order (LCS). However, more research is needed to improve the way we measure cross-modal coordination. The LCS measures we used, are still too simple to accurately capture ordering effects, especially when sentence processing is in action (i.e., encoding and production): as only one subsequence, out of many possible subsequences, is retained, and the temporal dimension (e.g., fixation duration or word length) over which such alignments take place is not yet fully accounted.

We also observed that the correlation between scan patterns and sentences varies between task phases: our mixed effects model showed a significant main effect of Region, as well as an interaction between Region and sentence similarity. This can be explained by the fact that the role of cross-modal coordination varies between the three task phases. During Planning, the visual system performs a search for the cued target object and is strongly guided by both target properties and referential scene information (Malcolm & Henderson, 2010; Castelhano & Heaven, 2010). During Encoding, once the target object has been identified, linguistic processing draws upon visual processing in order to retrieve material to be used in production: information about the identity, locations, and properties of objects. Finally, during Production, the objects in the scene are re-inspected by the visual system; crucially, this happens in the order in which they are mentioned, as fixations on objects are known to precede mentions by a fixed interval (Griffin & Bock, 2000).

The fact that we consistently find cross-modal coordination across different phases suggests that the coordination of visual and linguistic processing is not limited to the process of overt language production per se. Our data shows that scan patterns in Planning and Encoding are already predictive of the sentence that the participant will produce, before they start speaking. Visual attention can therefore be thought of as constructing the referential link between sampled visual information and the linearization of words in a description. This happens in preparation for speaking, and therefore the coordination of scan patterns and sentences precedes the onset of speech.

Our results also show that the selection of referents is modulated by both visual and linguistic mechanisms. As originally observed by Gleitman et al. (2007), we confirm that orienting visual attention towards a particular visual referent has an influence on the sentences that are generated. Moreover, we demonstrate that the semantic properties of the cued referent, e.g., animacy, influence the degree of cross-modal coordination observed, and interact with the effect of the phase of the task. Nevertheless, the range of factors and mechanisms modulating the dynamics of cross-modal coordination have yet to be fully explored.

The fact that similar scan patterns are associated with similar sentences during language production means that we can predict what participants say based on where they look. More specifically, we can predict which objects participants will mention, and in which order. Not only the objects themselves, also the relationships between objects can be deduced from scan patterns. For

example the fixation sequence man-L, arm-L, paper could be indicative of the utterance *the man writes on the paper* in Fig. 1. Even though the verb *write* does not correspond directly to a depicted object, its mention can be predicted if sequences similar to man-L, arm-L, paper tend to co-occur with the word *write*.

To illustrate the predictive property of scan patterns, we developed an algorithm which uses scan pattern similarity to retrieve the sentence associated with a given scan pattern from a pre-existing pool of sentences. Our algorithm predicts the correct sentence better than chance, and returns a sentence that is similar to the target if it fails to retrieve the exact target sentence. This result confirms the theoretical validity of our cross-modal coordination hypothesis and shows how it can be applied for sentence prediction.

Predicting what people will say based on their scan patterns has practical applications. For example Qu and Chai (2007) use gaze information to improve speech recognition; their approach works because of the effect demonstrated in this paper: recognizing a word is easier if we know which object is fixated when the word is spoken. In related work, Prasov and Chai (2010) show that co-reference resolution can also benefit from the availability of gaze information: resolving a pronoun such as *it* is facilitated if the system knows which objects are fixated concurrently. Furthermore, a number of studies have demonstrated that gaze information is useful for vocabulary acquisition, both by humans and by computers (Frank, Goodman, & Tenenbaum, 2009; Yu & Ballard, 2007); also this finding is a consequence of the fact that similar scan patterns occur with similar sentence productions.

## References


Andersson, R., Ferreira, F., & Henderson, J. M. (2011). The integration of complex speech and scenes during language comprehension. *Acta Psychologica*, *137*, 208–216.

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

Burnard, L. (1995). Users guide for the British National Corpus [Computer software manual]. British National Corpus Consortium, Oxford University Computing Service.

Castelhano, M., & Heaven, C. (2010). The relative contribution of scene context and target features to visual search in real-world scenes. *Attention, Perception and Psychophysics*, *72*, 1283–1297.

Coco, M. I., & Keller, F. (2010). Sentence production in naturalistic scene with referential ambiguity. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1070–1075). Portland, OR: Cognitive Science Society.

Cristino, F., Mathot, S., Theeuwes, J., & Gilchrist, I. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behaviour Research Methods*, *42*, 692–700.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (2003). *Biological sequence analysis: Probabilistic models of proteins and nucleid acids*. Cambridge,UK: Cambridge University Press.

Findlay, J. M., & Gilchrist, I. D. (2001). Visual attention: The active vision perspective. In *M. Jenkins & L. Harris (Eds.), Vision and attention* (pp. 83–103). New York: Springer-Verlag.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? spatial and sequential aspect of fixations during encoding and recognition. *Journal of Vision*, *8*(2), 1–17.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Gleitman, L., David January, D., Rebecca Nappa, R., & Trueswell, J. (2007). On the give and take between event apprehension and utterance formulation. *Journal of Memory and Language*, *57*, 544–569.

Gomez, C., & Valls, A. (2009). A similarity measure for sequences of categorical data based on the ordering of common elements. *Lecture Notes in Computer Science*, *5285/2009*, 134-145.

Griffin, Z., & Bock, K. (2000). What the eyes say about speaking. *Psychological science*, *11*, 274-279.

Gusfield, D. (1997). *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge,UK: Cambridge University Press.

Hayhoe, M. (2000). Vision using routines: a functional account of vision. *Visual Cognition*, *7*, 43-64.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*(4), 219-222.

Howell, D. C. (2002). *Statistical methods for psychology*. Pacific Grove, CA, USA: Wadsworth Group, Duxbury, Thomson Learning.

Humphrey, K., & Underwood, G. (2008). Fixation sequences in imagery and in recognition during the processing of pictures of real-world scenes. *Journal of Eye Movement Research*, *2*(2), 1–15.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(1), 1489–1506.

Knoeferle, P., & Crocker, M. W. (2006). The coordinated interplay of scene, utterance and world knowledge. *Cognitive Science*, *30*, 481–529.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, *25*, 296–324.

Land, M. F., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*, 13111328.

Landauer, T., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284.

Malcolm, L., G., & Henderson, J. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, *10(2)*(4), 1–11.

Mitchell, J., & Lapata, M. (2009). Language models based on semantic composition. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 430- 439.

Neider, B., M., & Zelinsky, G. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*, 614–621.

Potter, M. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 509-522.

Prasov, Z., & Chai, J. Y. (2010). Fusing eye gaze with speech recognition hypotheses to resolve exophoric references in situated dialogue. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 471–481). Cambridge, MA: Association for Computational Linguistics.

Qu, S., & Chai, J. (2007). An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 284–291). Rochester: Association for Computational Linguistics.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*(1-3), 151-173.

Schmidt, J., & Zelinsky, G. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, *62*(10), 1904-1914.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*(268), 632–634.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, *4*(113), 766–786.

Vo, M., & Henderson, J. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, *10*(3), 1–13.

Yang, H., & Zelinsky, G. (2009). Visual search is guided to categorically-defined targets. *Vision Research*, *49*, 2095–2103.

Yu, C., & Ballard, D. (2007). A unified model of word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–2165.

## Appendix A
## Validation against Additional Similarity Measures

In this supplementary material, we confirm the validity of our cross-modal coordination hypothesis by demonstrating that our main result (scan pattern similarity increases with increasing linguistic similarity) holds when alternative similarity measures are employed. First, we discuss in detail how the LCS measures, which are partially sensitive to order, compare with Bag Of Words (BoW) measures, which do not take order into account. Then, we show that cross-modal coordination is obtained even when time information is included in the scan pattern representation, and also when lexical semantics based on vector composition is used to measure the similarity of sentences.

Furthermore, we provide a more detailed explanation of the procedure adopted to select the linear mixed effect models.

***BoW Vs LCS.***

The LCS is a measure of similarity based on common sub-sequences. A sub-sequence retains the ordering of common elements that two sequences share. Thus, in order to distinguish the role played by ordering from that of common elements when computing similarities of sentences and scan patterns, we compare LCS with a measure purely based on common elements (i.e., irrespective of the order in which they occur). We use BoW, which counts the number of common elements between two sequences relative to their total number of elements. We calculate BoW on both scan patterns and sentences. Then, we test the difference of strength on the correlation coefficients between LCS measures (LCS.L, LCS.V) and BoW measures. Using the Fisher $\rho - to - z$ transformation, we calculated the value of $z$ to assess the significance of the difference between the two correlation coefficients $\rho_{\mathrm{LCS}}$ and $\rho_{\mathrm{BoW}}$. After performing a two-tailed test at $\alpha = 0.05$ on $z$, we analyze the $p - value$ (refer to Howell, 2002, pp. 277-278).

The difference between the correlation coefficients depends on the sample size. For this reason, we perform the significance test across samples of increasing size. This gives us a detailed comparison between BoW and LCS. Our aim is to show that already small sample sizes yield a significant difference between BoW and LCS.

Fig. A1 plots $p$-value as a function of sample size for the three regions of analysis. We find that already at a sample of size of 6000, the difference in the coefficients is significant for all the three regions. When looking at the individual regions, we observe that for Planning significance is reached for sample size as small as 250. Encoding and especially Production need larger sample sizes to stabilize. This analysis confirms that ordering is important to capture cross-modal coordination. However, the difference in effect size (correlation coefficients) is small; as discussed in the main paper, we believe that LCS is sub-optimal in capturing the pattern of similarity. The main drawback is that LCS returns only one among all possible ordered sub-sequences, de facto discarding the information of all other sub-sequences. Also, LCS does not capture efficiently temporal information, which is crucial when visual and linguistic processing have to be synchronize, i.e., during encoding and production. In the next paragraph, we show that including temporal information on the scan pattern could help improving the correlation during these phases.

***Alternative Measures.***

As alternative measure for similarity between sequences of categorical data, we adopt Ordered Sequence Similarity (OSS, Gomez & Valls, 2009), which can be used to quantify similarity both for scan patterns and word sequences and has been shown to be more effective than simpler measures such as edit distance. In addition, we employ a similarity measure based on Latent Semantic Analysis (LSA, Landauer, Foltz, & Laham, 1998) to compare word sequences only.

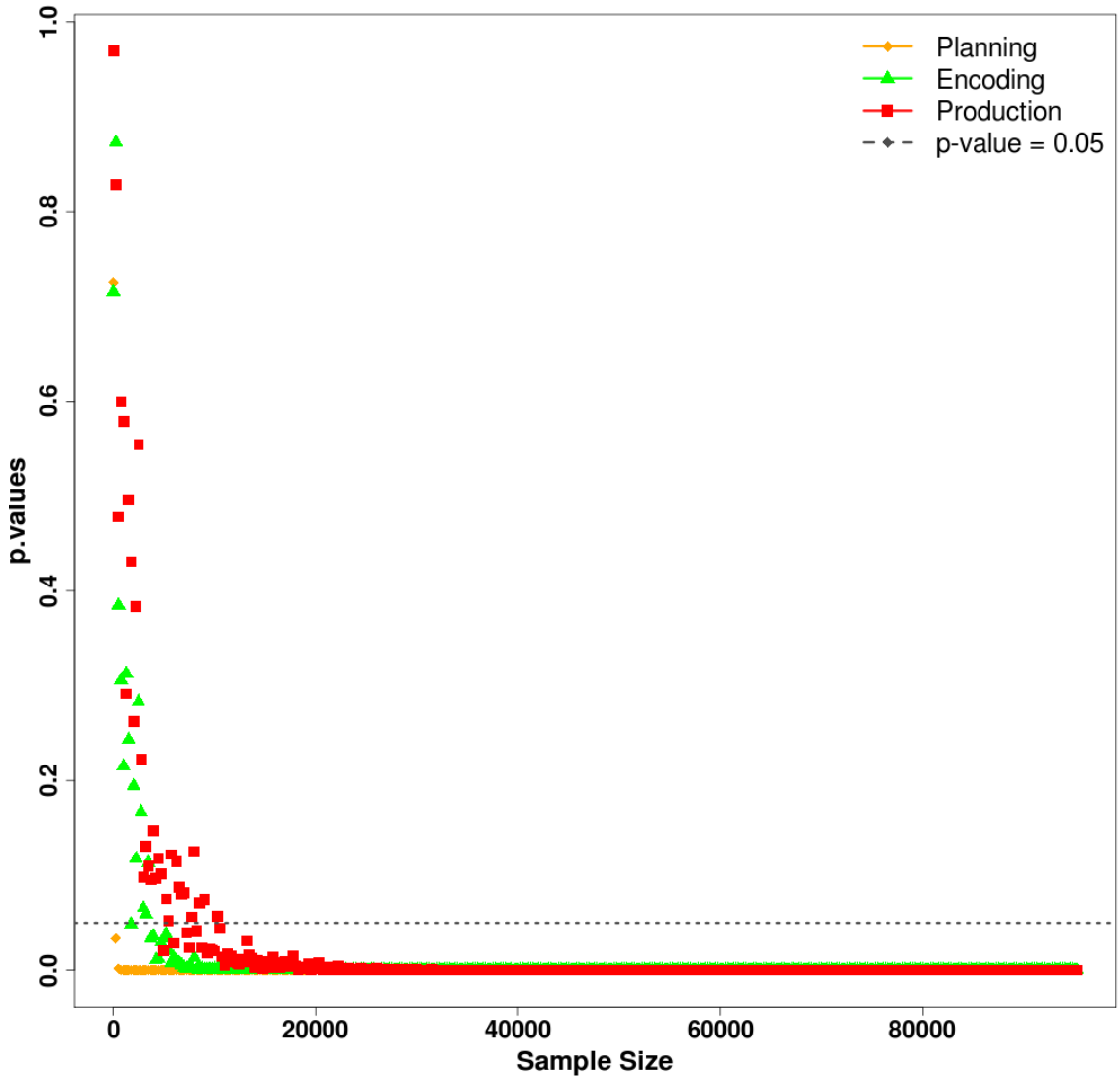



*Figure A1*. Test of significance for difference between correlation coefficients BoW and LCS, for random samples of increasing size (from 10 to whole dataset in steps of 250), across the three regions of analysis (Planning, Encoding and Production).

OSS is a measure based on two aspects of sequential data: the elements a sequence is composed of and their positions. When comparing two sequences, OSS divides the elements in common (shared) and uncommon (unique) ones; on the shared elements, it takes into account their relative position. For example in Fig. 2 in the main article, four objects are shared by the two scan patterns (man-L, plant, statue, suitcase) and two objects (telephone, man-R) are unique to SP1 and SP2, respectively. For each common element, the distance between the two sequences is calculated, e.g., statue in SP1 is one unit away from statue in SP2. Distances between common elements and the number of uncommon elements are combined into a single score and normalized based on sequence length (for details refer to Gomez & Valls, 2009). Despite its name, OSS is a dissimilarity measure, with values ranging from 0 for most similar to 1 for least similar. To facilitate comparison with the other measures, we convert OSS values to similarity values by subtracting them from 1.

We also experimented with OSS computed over scan patterns that include time information (OSS-Time): for this measure, we relabeled the objects by including an index indicating temporal progression into the label. This index was obtained by dividing the fixation duration on a given object into 50 ms intervals. For example, if the object man was fixated for 150 ms, then we divided the fixation on this object in three slots of 50 ms each and assigned the labels man-1, man-2, and man-3.

Latent Semantic Analysis (LSA, Landauer et al., 1998) is a widely used computational model of word meaning. LSA measures the similarity between words based on the co-occurrence of context words within the same document. Intuitively, two words are semantically similar if they occur in similar contexts. LSA represents words as vectors of co-occurrence counts, and semantic similarity is quantified as vector distance. For this study, we used a version of LSA that goes beyond word-level representations by computing LSA vectors for sentences (Mitchell & Lapata, 2009). In this approach, the meaning of a sentence is represented as the composition of the vectors of the words in the sentence. We built our LSA model using the British National Corpus, which contains 100 million word of text and speech (Burnard, 1995). For each sentence in our data set, we computed an LSA vector for each content word in the sentence (context window of size five; low frequency words were removed). We then combined these vectors using addition to obtain sentence vectors (an alternative discussed by Mitchell & Lapata, 2009, would be vector multiplication). Similarity between sentence vectors was measured using cosine distance. As LSA is a linguistic measure, it can be computed only on sentences.

***Results.***

Fig. A2 plots linguistic similarity (LCS.L and LSA) as a function of scan pattern similarity incorporating time information (OSS-Time). The figure shows that scan pattern similarity increases with increasing sentence similarity. This confirms the results in the main text, where we found the same pattern with only LCS (LCS.L increases as a function of LCS.V). We can therefore conclude that this result generalizes to other similarity measures (OSS and LSA) and is valid also when we include temporal information (fixation duration) in scan pattern similarity (as in OSS-Time).[4]

Tab A1 presents the results of correlations analyses involving all pairs of similarity measures. When time is included in the scan pattern similarity measure (OSS-Time) and paired with the sequential linguistic similarity measure (LCS.L), we find a significant positive correlation with a similar coefficient across all three regions of analysis. This replicates the results for the pairing of LCS.V and LCS.L in the main article.

[4]OSS without time (not reported here) shows the same trend.

Table A1
*Correlations (Spearman ρ) between the different similarity measures across regions of analysis: Planning, Encoding, and Production. All correlations are significant at $p < 0.05$, with the exception of the correlation of OSS-Time and LSA in the Planning region.*

| Measures | LCS.V | | | OSS-Time | | | LSA | | |
|---|---|---|---|---|---|---|---|---|---|
| | Plan | Enc | Prod | Plan | Enc | Prod | Plan | Enc | Prod |
| OSS-Time | 0.82 | 0.80 | 0.77 | | | | | | |
| LSA | 0.10 | 0.11 | 0.13 | 0 | 0.11 | 0.22 | | | |
| LCS.L | 0.48 | 0.47 | 0.38 | 0.34 | 0.39 | 0.35 | 0.36 | 0.35 | 0.38 |

When pairing LCS.V and OSS-Time with LSA, our alternative measure of sentence similarity, we obtained a significant correlation in all but one case (OSS-Time and LSA during Planning); see Fig. A3 for a scatter plot and LME estimates. However, the correlation coefficients obtained are lower (with ρ ranging from 0.10 to 0.22) than if we compare LCS.V and OSS-Time with LCS.L (with ρ ranging form 0.34 to 0.48). LSA is based on word co-occurrences; it represents the lexical meaning of a sentence, rather than its word order (it is a non-sequential measure). LCS.V focuses on the similarity of word sequences, disregarding any semantic relationships between words, and is therefore expected to be more highly associated with sequential scan pattern measures, which measure the similarity between sequences of fixated objects.

Note the correlation between the two linguistic similarity measures LCS.L and LSA is relatively weak (with ρ ranging from 0.35 to 0.38). This also confirms that these measures to some extent measure different things. As expected, the correlation between the two sequential measures LCS.V and OSS-Time is high (with ρ ranging from 0.77 to 0.82). The results of the correlation analysis are confirmed in the mixed effect analysis, see Tab A2 for full list of coefficients.

In line with the results observed at the global level, the sequential similarity measures (LCS.L/OSS-Time) achieve a higher correlation coefficients than the other measures also on the local level; see Tab A3.

An important point emerging from this analysis regards the role of temporal information with respect to the phases of the task. We find that when temporal information is included in the scan patterns, both Encoding and Production have a significantly higher similarity than Planning (main effects). Moreover, when OSS-Time is paired with LSA, we find a positive interaction of LSA with the Production region. This positive interaction is not found with any other combination of measures. This might suggest that temporal information is implicitly accounted for by statistical properties of LSA, and when paired with a visual measure that explicitly includes temporal information (OSS-Time), we could better capture changes in cross-modal coordination during Production.

The measures presented in this paper, however, are limited in dealing with time; thus in future work, we are planning to develop measures of similarity which integrate and evaluate the effect of time in both sentence and scan pattern, thus making possible to capture more closely the dynamics of coordination allowing visual and linguistic processing to be synchronized.

In summary, our supplementary analyses broadly confirm the results presented in the main text: scan pattern similarity increases with increasing linguistic similarity. The result holds across a range of different similarity measures, and even when fixation duration is considered. The effects is

Table A2

*Coefficients for the mixed effects model analysis. The dependent measure is scan pattern similarity (LCS.V or OSS-Time); predictors are sentence similarity (LCS.L or LSA) and Region (contrast coded with Embedding as the reference level) and Cue (contrast coded with Inanimate as the reference level. Note that the minimal models are shown; a factor marked "n.i." has not been included during model selection.*

| Predictor | LCS.L/LCS.V | LCS.L/OSS-Time | LSA/LCS.V | LSA/OSS-Time |
|---|---|---|---|---|
| (Intercept) | $0.188^{***}$ | $0.428^{***}$ | $0.193^{***}$ | $0.430^{***}$ |
| Sentence | $0.514^{***}$ | $0.251^{***}$ | $0.206^{***}$ | $0.105^{***}$ |
| $\text{Region}_{\text{PlanVsEnc}}$ | $0.038^{***}$ | $-0.067^{***}$ | $0.035^{***}$ | $-0.068^{***}$ |
| $\text{Region}_{\text{ProdVsEnc}}$ | $-0.018^{***}$ | $0.063^{***}$ | $-0.016^{***}$ | $0.063^{***}$ |
| $\text{Cue}_{\text{AniVsIna}}$ | $0.038^{***}$ | $-0.004^{***}$ | $0.008^{***}$ | *n.i.* |
| $\text{Cue}_{\text{MixVsIna}}$ | $0.005^{***}$ | $0.002^{***}$ | *n.i.* | *n.i.* |
| Sentence:$\text{Region}_{\text{ProdVsEnc}}$ | $-0.352^{***}$ | $-0.140^{***}$ | $-0.100^{***}$ | $0.051^{***}$ |
| Sentence:$\text{Region}_{\text{PlanVsEnc}}$ | $0.326^{***}$ | $0.069^{***}$ | $0.091^{***}$ | $-0.056^{***}$ |
| Sentence:$\text{Cue}_{\text{AniVsIna}}$ | $0.093^{***}$ | $0.044^{***}$ | $-0.029^{***}$ | *n.i.* |
| $\text{Cue}_{\text{AniVsIna}}$:$\text{Region}_{\text{ProdVsEnc}}$ | $-0.019^{***}$ | *n.i.* | $0.013^{***}$ | *n.i.* |
| $\text{Cue}_{\text{AniVsIna}}$:$\text{Region}_{\text{PlanVsEnc}}$ | $0.008^{***}$ | $0.007^{***}$ | *n.i.* | *n.i.* |
| $\text{Cue}_{\text{MixVsIna}}$:$\text{Region}_{\text{ProdVsEnc}}$ | $0.014^{***}$ | $0.006^{***}$ | *n.i.* | *n.i.* |
| $\text{Cue}_{\text{MixVsIna}}$:$\text{Region}_{\text{PlanVsEnc}}$ | *n.i.* | $-0.002^{*}$ | *n.i.* | *n.i.* |

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

found in within and the between scene analysis, and across the different phases of the task.

## Appendix B
## Model Selection Procedure

The mixed effects were fitted using maximum likelihood estimation. All fixed factors were centered to reduce collinearity; four random factors were considered: participant 1, participant 2, trial 1, and trial 2; note that the data comes in pairs, which is why two random variables for each of participant and trial are necessary.

The minimal model was selected by following a forward step-wise procedure which compares nested models based on log-likelihood model fit. We start with an empty model, to which we first add the random factors. Once all random factors have been evaluated as intercepts, we proceed by adding the fixed factors. For every fixed factor, we calculate whether the fit of the model improves when random slopes for that factor are included. By including random slopes, we take into account the variability of each fixed effect (e.g., scan pattern similarity) for the different grouping levels of the random effects, e.g., participants. To ensure computational tractability, we consider random slopes only on the main effects, i.e., not for interactions. Factors are added one at time, and ordered by the amount of improvement in model fit they achieve. Only factors (fixed or random) that significantly improve model fit are added.

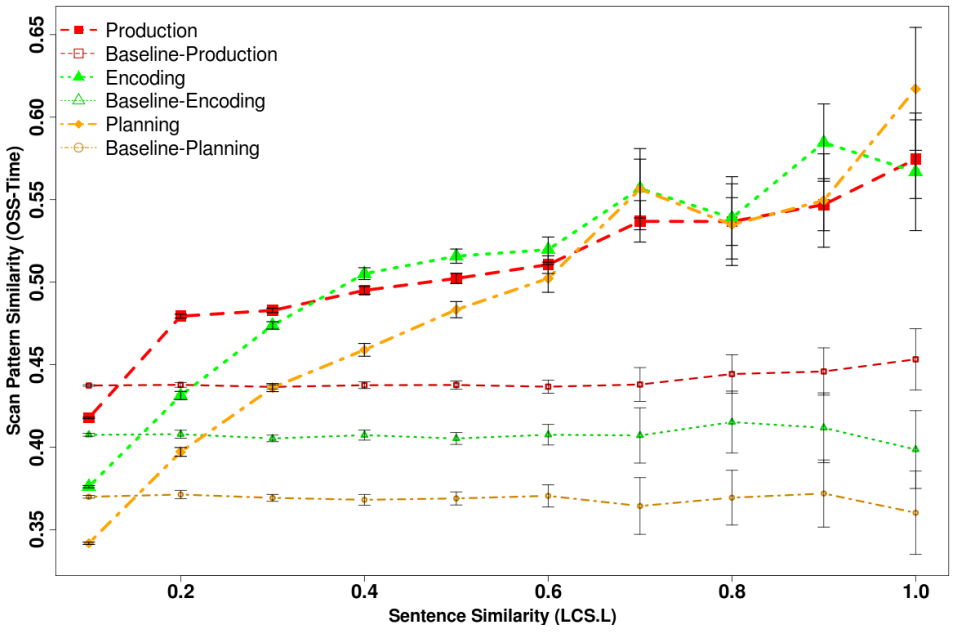
We test for a significant improvement in model fit using a $\chi^2$ test that compares the log-likelihood of the model before and after adding the new factor. Overall, this procedure returns a

Table A3

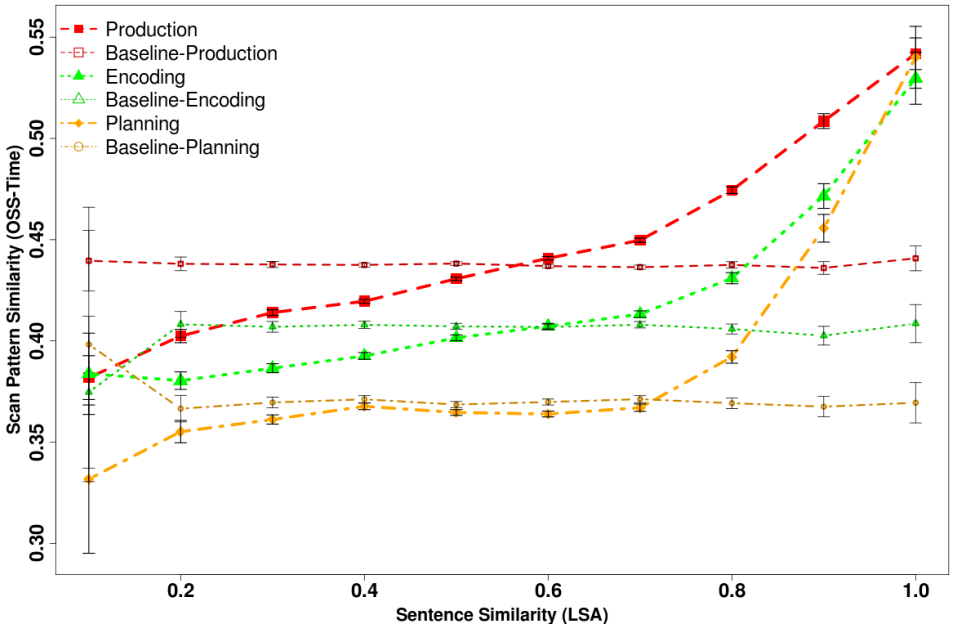
*Minimum and maximum correlations (Spearman ρ) between the different similarity measures across regions of analysis (Planning, Encoding, Production); correlations were computed for each scene separately. All correlations are significant at* $p < 0.05$.

| Measures | | LCS.V | | | OSS-Time | | | LSA | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Plan | Enc | Prod | Plan | Enc | Prod | Plan | Enc | Prod |
| OSS-Time | Min | 0.73 | 0.57 | 0.39 | | | | | | |
| | Max | 0.88 | 0.89 | 0.91 | | | | | | |
| LSA | Min | −0.06 | −0.12 | −0.08 | −0.36 | −0.28 | −0.1 | | | |
| | Max | 0.29 | 0.35 | 0.26 | 0.15 | 0.35 | 0.36 | | | |
| LCS.L | Min | 0.25 | 0.14 | 0.04 | −0.16 | 0.06 | −0.1 | 0.19 | 0.14 | 0.18 |
| | Max | 0.63 | 0.77 | 0.51 | 0.52 | 0.65 | 0.51 | 0.57 | 0.65 | 0.52 |

model that maximizes fit with the minimal number of predictors.




(a)




(b)

*Figure A2*. Scan pattern similarity (OSS-Time) as a function of linguistic similarity (LCS.L and LSA) across all 24 scenes. The data has been binned on the x-axis; the whiskers represent 95% confidence intervals. The baselines were obtained by randomly pairing a sentence and a scan pattern from the respective phase.

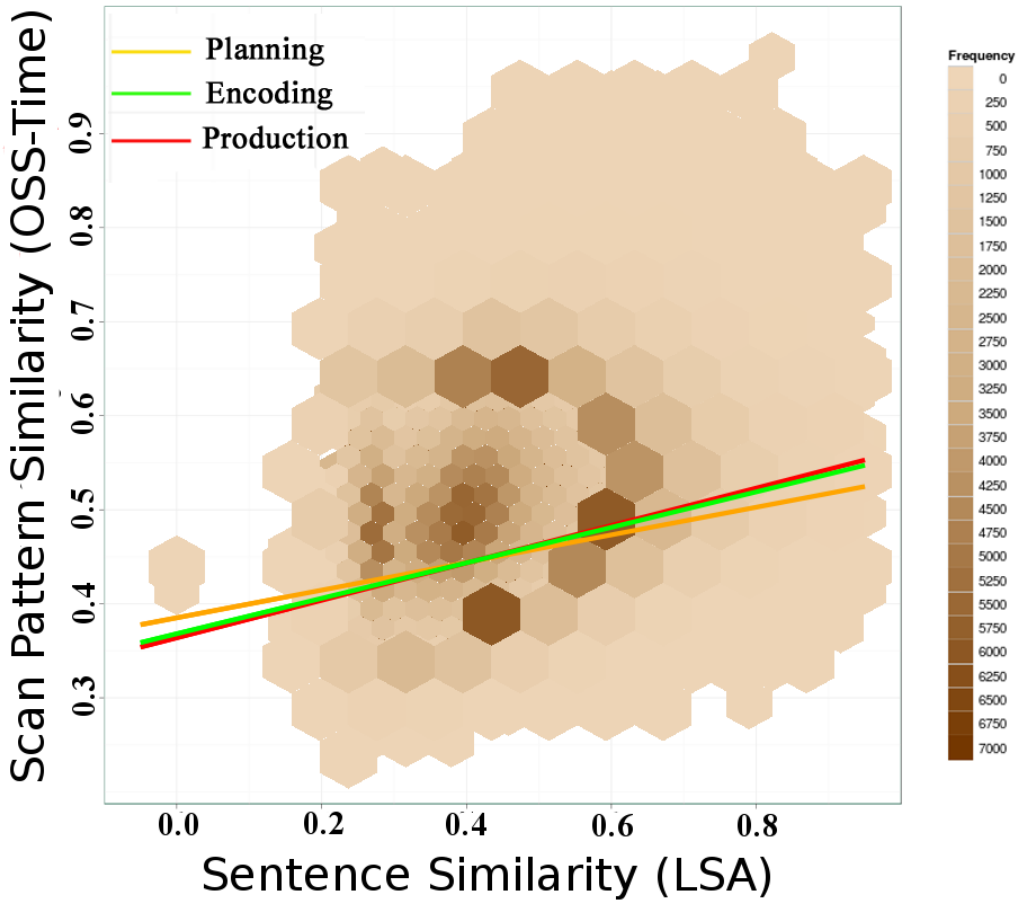



*Figure A3*. Hexagonal plot of predicted values of the linear mixed effects model: linguistic similarity predicted by scan pattern similarity. The plot shows the observed data binned into hexagons. The color of the hexagon reflects the frequency of the observations within it (darker for more observations). The solid lines represent the grand mean intercept and slope: Planning (orange), Encoding (green), Production (red).