Alignment and Task Success in Spoken Dialogue

David Reitter[a], Johanna D. Moore[b]

[a] *The Pennsylvania State University*
[b] *University of Edinburgh*

Abstract

Task-solving in dialogue depends on the convergence of the situation models held by the dialogue partners. The Interactive Alignment Model (Pickering and Garrod, 2004) suggests that this convergence is the result of an interactive alignment process, which is based on mechanistic repetition at a number of linguistic levels. In this paper, we develop two predictions arising from the theory, along with two methods to quantify the known structural priming effects in the full inventory of syntactic choices found in text and speech corpora. (a) Under a rational perspective, we expect increased repetition in task-oriented dialogue compared to spontaneous conversation. We find within- and between-speaker priming in a corpus of spontaneous conversations, but stronger priming in task-oriented dialogue. (b) The Interactive Alignment Model predicts linguistic adaptation to be correlated with task success. We show this effect in a corpus of task-oriented dialogue, where we find a positive correlation of long-term adaptation and a quantifiable task success measure. We argue that the repetition tendency relevant for the high-level alignment of situation models is based on slow adaptation rather than short-term priming. We demonstrate that lexical and syntactic repetition are reliable and computationally exploitable predictors of task success.

Keywords: syntactic priming, structural priming, task success, interactive alignment, dialogue, computational psycholinguistics

Introduction

Humans appear to be remarkably efficient communicators in light of the computational complexity of natural language. Dialogue poses many challenges: interlocutors have different viewpoints, linguistic preferences and knowledge states. What may help is that we are copy cats rather than creators; we prefer to adapt our language rather than to go against the grain. The *Interactive Alignment Model* (IAM, Pickering and Garrod, 2004) posits that such mutual adaptation is easier than careful selection of information and targeting of the message in dialogue. The IAM suggests that basic priming effects at lower processing levels (lexical, syntactic) reinforce alignment at higher ones (e.g., semantic, pragmatic), leading to linguistic adaptation and grounding of situation models during speaker interaction. Priming occurs when memory retrieval is biased by previous context; in this case, priming refers to a tendency to choose linguistic constructions that have been used shortly beforehand.

The IAM assumes that this repetition of linguistic choices is not just an artifact of general memory retrieval properties, but instead is a mechanism (alignment) by which interlocutors build a common understanding of the situation, enabling them to successfully communicate without keeping track of one another's linguistic idiosyncrasies. According to the IAM, repetition is a heuristic that helps establish common ground unless the situation requires more careful monitoring and modeling of one's interlocutor's state of knowledge.

The success of our interactions varies. The success of task-oriented dialogue depends on communication and is quantifiable, allowing us to test the IAM by linking it to alignment. In this paper, we correlate priming at levels of sentence structure (syntax) and word choice, the problem-solving objective of the dialogue, and success.

Hypotheses

Humans align their linguistic choices at several representational levels. At a low level, phonetic reductions occur in jointly understood words (Bard et al., 2000). An example of adaptation at a higher level of representation involves dialogue partners that develop coherent situation models, as in Garrod and Andersons's (1987) *Maze Game* study. The task was designed to elicit a coordinated communication system between participants. They found that speakers tended to make the same semantic and pragmatic choices as in the utterances they had just heard. As the games proceeded, participants developed a common description scheme for positions in the maze.

However, the full causal cascade from lower-level priming to high-level alignment has not yet been observed. Specifically, the hypothesized correlation between the two, and ultimately successful communication, has eluded empirical verification.

In this paper, we focus on implicit linguistic decisions: the basic mechanics of communication implemented in syntactic structure, as opposed to the high-level strategies speakers use to describe aspects of a task, or the more explicitly controlled lexical choices. *Syntactic priming* occurs when speakers show a tendency to prefer one phrase structure over an available alternative shortly after having used this structure or having heard an interlocutor use it (Bock, 1986). Verbatim, lexical repetition is known to increase the strength of priming (Pickering and Branigan, 1998; Gries, 2005; Hartsuiker et al., 2008). This *lexical boost* is a crucial effect for the IAM, as it shows propagation of alignment from lower to higher levels of representation.

Thus far, there is only limited evidence for the occurrence of structural adaptation outside of carefully controlled laboratory settings. As we will see, speakers also adapt in situated, realistic dialogue. For example, consider this excerpt from the Map Task corpus (Anderson et al., 1991; McKelvie, 1998), a dataset that we will use extensively in this study. One speaker (g) is giving directions for another one (f) to follow on a map:

f: *from the mill wheel and up to the abandoned cottage to the right* ***like a tick shape*** *it'd be s–* ***[the shape of a tick]*** *from the the*
g: *no*
g: ***[the shape of a] [like an oval shape]*** *from the caravan park you start just above the caravans*

Here, g first sets out to repeat the latest syntactic construction (*the shape of an oval*), but proceeds to use an alternative one (*like an oval shape*) in its repair, mirroring his interlocutor's first syntactic choice (*like a tick shape*). The spontaneous syntactic choice is a direct repetition, but would be ungrammatical if completed (*the shape of a oval*). Both of g's expressions reflect structural repetitions rather than plausible alternatives to describe an oval-shaped path. This example of repetition reflects not only syntactic, but also lexical choices. A quantitative model of priming should cover such cases, but also repetitions that occur outside of lexically or semantically similar contexts. In our study, we are concerned with implicit (syntactic) effects. We therefore measure priming of syntactic phrase-structure rules, whereby word-by-word repetition (topicality effects, parroting) is explicitly excluded.

We examine the IAM from a functional perspective, and derive two groups of testable hypotheses. The first examines *syntactic priming in task-oriented dialogue*, while the second adds a functional perspective by showing a *correlation between adaptation and task success*.

Our first hypothesis concerns the mechanisms of priming. Syntactic priming is claimed to be a mechanistic effect, though this does not necessarily mean that it is automatic and agnostic to contextual influence. According to some cognitive architectures (Anderson and Lebiere, 1998), priming effects are the result of working memory activity. From a functional and rationalist point of view, the enhancement of communication by priming suggested by the IAM could have led to an architectural configuration where the demands of the dialogue situation drive syntactic priming. For instance, syntactic representations may be temporarily associated with semantic ones. Topics determine semantics held in working memory, and so, meaning is typically clustered rather than randomly mixed. In line with this, theories of dialogue have suggested clustering of topics, and coherence of topic structure (Grosz and Sidner, 1986; Grosz et al., 1995). Given any syntactic-semantic associations, syntactic structure may tend to cluster as well.

We hypothesize that there is a tendency for dialogue partners to repeat syntactic structure within brief time windows, and that they do more so in task-oriented dialogue than in spontaneous conversation. Regardless of the underlying mechanisms, the IAM seems incompatible with the inverse hypothesis: less priming in task-oriented dialogue.

In the first set of experiments (1–2), we look at short-term priming effects and whether speakers implicitly use increased short-term adaptation in situations where they may benefit from it.

The second hypothesis is derived from the IAM's core idea connecting low-level priming to high-level mutual understanding and task success. Adaptation itself is difficult to manipulate in naturalistic human-human dialogue. However, we expect observable variation in adaptation levels.

The IAM predicts that task-oriented dialogues that exhibit more syntactic adaptation between the interaction partners will ultimately yield more task success. We test this prediction in Experiments 3–4. We conclude with an experiment that uses machine learning techniques to demonstrate that both syntactic and lexical alignment can be exploited to predict task success (Experiment 5).

We will refer to several different variants of syntactic adaptation. *Adaptation* denotes an increased amount of re-use of decisions compared to expected repetition occurring by chance. *Short-term priming* is short-lived adaptation, which disappears after a few seconds. *Long-term adaptation* is adaptation that is enhanced by repeated exposure, persistently increasing the availability of syntactic structures. *Alignment* is a cascade of adaptation processes between speakers at different linguistic levels postulated by the IAM. Alignment culminates in assimilated situation models and established ad-hoc conventions between speakers.

Interactive Alignment and Structural Priming in Dialogue

Structural priming is a special case of adaptation, either between or within speakers. Language production and comprehension are biased by recent experience, regardless of whether the structures were observed while comprehending language, or whether they were used in one's own speech. Alignment at the syntactic level is well-documented and known to occur in a variety of contexts: between questions and answers (Levelt and Kelter, 1982), in comprehension and production. It can be specific to dialogue partners (Brennan and Hanna, 2009) or to the perceived abilities of an interlocutor (Branigan et al., 2011).

Bock (1986) established the experimental paradigm that uncovered structural priming in speech. Bock and Loebell (1990) demonstrated evidence for priming of syntactic structure independent of semantics and metrical or event structure. Pickering and Branigan (1998) found syntactic priming in written language production using scripted situations and a sentence completion task. Branigan et al. (2000) found clear evidence for syntactic alignment in dialogue-like lab interactions. Their experimental design is prototypical of much of the experimental work in structural priming. In their experiments, dialogue partners took turns describing pictures to one another to enable their partner to identify the card containing the described picture from a set of cards laid out in front of them. One of the speakers was a confederate and produced descriptions based on a script that manipulated syntactic choice, in particular whether a double object or a prepositional object construction was used (e.g., *the cowboy giving the clown a balloon* vs. *the cowboy giving a balloon to the clown*). The syntactic structure of the confederate's description strongly influenced the syntactic structure of the subject's description in the turn immediately following.

Two adaptation effects occur: (a) fast, short-term and short-lived priming, and (b) slow, long-term adaptation that persists and is likely to be a result of implicit learning (see Ferreira and Bock (2006); Pickering and Ferreira (2008) for reviews). Long-term adaptation is a learning effect that can persist over several days (Bock et al., 2007; Kaschak et al., 2011b). Recent work has proposed models that explain the mechanisms of the effects (Bock and Griffin, 2000; Kaschak et al., 2011a) within the context of language acquisition (Chang et al., 2006) and eneral memory retrieval (Reitter et al., 2011). The remainder of this article will address short-term syntactic priming first, and then discuss experiments with long-term syntactic and lexical alignment.

Most of the results on priming and alignment come from controlled experiments. We caution that designs in which subjects do a task constructed to elicit linguistic target constructions many times may not be a true reflection of linguistic choices made by participants in natural, spontaneous real-life dialogue. For instance, findings regarding verb-argument preferences in experimental conditions do not always correlate well with corpus studies (Roland and Jurafsky, 2002). One reason why some linguistic laboratory experiments fail to faithfully reproduce real-world language use may be the complexity of linguistic choice as evidenced by models derived from corpora. Gries (2005) argues that experimental designs may effectively control only some confounds, but not the variety of factors that influence linguistic decision-making.

Such criticisms are addressed by work on language elicited outside of artificially created situations, often in the context of spoken dialogue (Levelt and Kelter, 1982; Estival, 1985; Bock and Kroch, 1989; Gries, 2005; Szmrecsanyi, 2006, 2005; Dubey et al., 2005). These studies corroborate the laboratory experiments and also show that structural priming occurs in spontaneously produced language. However, these studies employ a design pattern that contrasts the use of alternative syntactic choices sharing the same semantics (e.g., *She picks up the book* vs. *She picks the book up*). Typically, such use of explicit alternations limits corpus studies as well as lab experiments to a small set of predetermined syntactic rules or constructions, such as particle placement as in the example, active vs. passive voice, or double object (DO) vs. prepositional object (PO) use for arguments to verbs. This design also hinges on a very simple notion of semantics. One could object that active and passive constructions, for instance, are not semantically equivalent and carry different connotations and information statuses (Steedman, 2000). Syntactic alternations mark syntactic choice points, i.e., where a speaker must choose a construction to use. The corpus-based approach we follow refers to syntactic choices, but does not require alternations to define or even measure priming.

Pickering and Garrod (2004) argue that if the main reason that priming effects occur is to facilitate alignment, they will be particularly strong during natural interactions. Corpora provide an opportunity to quantify and contrast spontaneous processes and the interaction between linguistic choices and cognitive tasks. The next section will describe this methodology in detail.

Methodology: Measuring Short-Term Priming in Corpora

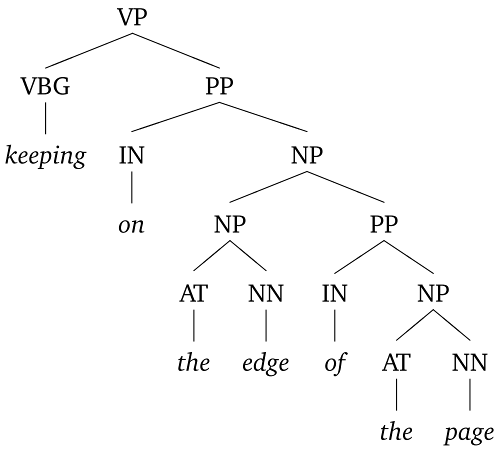
What we describe in the following is a method to quantify and contrast priming levels in datasets. They contain language spontaneously produced in contexts not designed to elicit syntactic priming or to test the IAM. The Switchboard corpus (Marcus et al., 1994) is a set of spontaneous telephone conversations; the HCRC Map Task corpus (Anderson et al., 1991) contains task-oriented dialogues.

Consider the following example. If we were to detect priming of passive constructions, we can do so with a range of different verbs and semantics by counting occurrences of passives, and contrasting the counts under two conditions: a repetition case (where a passive occurred shortly before), and a control case (where the passive has not occurred recently). Priming is the result of the difference between the normalized counts. Under this view, priming is not repetition, but the *increase in probability* caused by a preceding occurrence. Our technique is similar, but extends this method by looking at all syntactic constructions rather than just passives, and by using regression for greater sensitivity.

In this and other corpus studies, the concept of adding predictors as controls replaces the strict control of semantics in the laboratory experiment. We see a high degree of variance in speakers' choices of syntactic forms, which is natural, as the underlying semantics largely dictate how to construct the sentences. However, examining a large number of data points allows us to treat semantic variation as noise.

Corpus processing

To examine "all kinds of syntactic constructions", we analyze our datasets in terms of their syntactic phrase structure. Both of the corpora have been annotated with phrase structure trees through automatic and manual processes that included extensive verification (Marcus et al., 1994; Anderson et al., 1991). From the trees, we identify the syntactic rules used to construct them. We see the rules as a proxy for memory items that a speaker has to retrieve to produce or comprehend a sentence. For example, the tree



yields the six phrase structure rule instances shown in Table 1.[1]

The conversion from syntactic trees to rule instances is unambiguous.

Decay-based model of short-term priming

The amount of rule repetition can now be quantified. Structural priming predicts that a rule (*target*) occurs more often closely after a potential *prime* of the same rule (stimulus) than further away. Therefore, we correlate the probability of repetition with the distance between prime and target. For example, if a sentence-level conjunction leads to the rule $S \rightarrow S\ cc\ S$, and such a conjunction appears in utterances 3 and 11, we would observe a repetition, noting its distance $d = 8$ utterances. We sample repetitions and non-repetitions within 1-second or 1-utterance windows at different distances (ln(DIST), up to 25 utterances or 15 seconds). Thus, a rule occurrence in the dialogue will normally lead to up to 25 or 15 data points for the various distances, with a binary response variable indicating repetition vs. non-repetition. Memory effects generally decay nonlinearly. Analysis of the repetition probabilities over increasing d confirmed this distribution. ln(DIST) is therefore log-transformed in our models.

Unlike in controlled experimentation where specific syntactic constructions are elicited, every rule may be biased by a prior prime in this paradigm. The example shown in Figure 1 shows a subset of the rules appearing in the text. Repetitions α and β are both at distance 2, because the occurrences (prime and target) are two utterances apart, or 4.6 and 3.2 seconds, respectively. To facilitate the computation, we also drop all hapax rules (frequency $f = 1$).

We exclude cases where syntactic repetition is a mere consequence of verbatim lexical repetition (γ). The reason for this is that speakers may merely repeat such phrases without analyzing them syntactically. Lexical repetition is likely to result in syntactic repetition, which would possibly inflate results.

The basic statistical model compares the probability of a rule occurrence in situations when it was and wasn't primed. The null hypothesis is that this probability should be unaffected by the prime. Our statistical model is a sensitive variant of this idea. We predict the probability of repetition as a function of the time between prime and target. Priming effects decay over time or are subject to interference in working memory, so the effect assumes a decline of repetition probability with increasing distance between prime and target. The slope of this decline is the basis for comparison of priming strength under different conditions. The logistic regression model is specified in the appendix.

The effect of distance on syntactic repetition has been shown in related studies on corpora. Gries (2005) demonstrated a correlation of distance with the repetition probability of selected syntactic alternations in a corpus of spoken and written English. Gries found no effect of distances greater than one parsing unit (a unit similar to an utterance). Similarly, in our data, we see a strong decay only during the initial 5 seconds. In our method, unlike that of Gries, we take the distance effect on repetition within the short initial time period as a measure of short-term priming and determine how it interacts with other variables.

How repetition probability is modeled depends on assumptions about the underlying cognitive mechanisms. The first of two common views, *temporal decay*, implies a diminishing of repetition probability or priming effects over time. This assumes a form of decision-making that is influenced by decaying activation. The alternative view assumes *interference* of other material, resulting in a similar reduction in repetition probability. In this case, the selection of syntactic rules is influenced by interference from more recent syntactic structures even if they are inappropriate in light of contextual or semantic constraints (see Jonides et al., 2008, for a review of the two views of short-term memory). The latter may also suggest the influence of working memory on syntactic decisions, where working memory provides cues that aid in retrieval of memory. Short-term priming can be modeled as a combination of rapid temporal decay of syntactic information, and cue-based memory retrieval subject to interfering and facilitating semantic and other information in working memory (Reitter et al., 2011). The interaction of multiple activation mechanisms is a common assumption of ACT-R (Anderson and Lebiere, 1998). An additional difficulty in modeling such phenomena arises from the fact that one mechanism (e.g., temporal) may act as a proxy for the other (e.g., interference-based). So, although the rational analysis of memory retrieval needed in typical environments or text corpora may suggest temporal decay at the computational level (Marr, 1982), the underlying cognitive processes and neural implementation may be different (Lewandowsky et al., 2004).

The initial experiment 1 models distance between stimulus

[1] The analysis uses the Brown Corpus part-of-speech tags Kucera and Francis (1967). IN: preposition, AT: determiner, VBG: verb, present participle/gerund. CC: sentence-level coordinating conjunction.

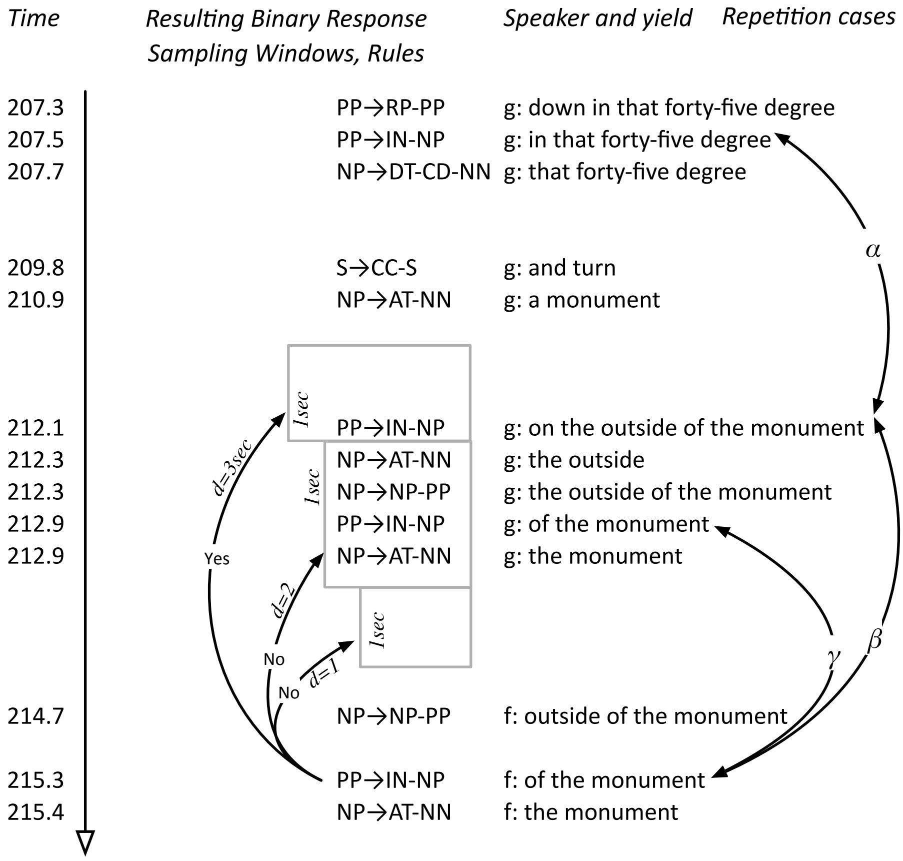

Time
Resulting Binary Response
Sampling Windows, Rules
Speaker and yield
Repetition cases
207.3 PP→RP-PP g: down in that forty-five degree
207.5 PP→IN-NP g: in that forty-five degree
207.7 NP→DT-CD-NN g: that forty-five degree
209.8 S→CC-S g: and turn
210.9 NP→AT-NN g: a monument
212.1 PP→IN-NP g: on the outside of the monument
212.3 NP→AT-NN g: the outside
212.3 NP→NP-PP g: the outside of the monument
212.9 PP→IN-NP g: of the monument
212.9 NP→AT-NN g: the monument
214.7 NP→NP-PP f: outside of the monument
215.3 PP→IN-NP f: of the monument
215.4 NP→AT-NN f: the monument
1sec
1sec
1sec
d=3sec
d=2
d=1
Yes
No
No
α
β
γ

Figure 1
Arrows on the right illustrate two instances of syntactic repetitions (α, β) and a lexical-syntactic one (γ) from Map Task. γ is not counted as it is also a lexical repetition. Arrows on the left show three samples (out of up to 15 per rule instance) connecting a rule instance of PP→IN NP (at bottom) with one-second time windows at varying distances d prior to the rule. The window at distance 3 contains repetition case β, yielding a positive sample (marked *"Yes"*). In the other two windows, there is no repetition, yielding negative samples.

Table 1
Syntactic rules and additional information extracted from the Map Task corpus. The speaker here is the direction follower, as opposed to the direction giver. This is a simplified example compared to the actual annotation.

Onset time (s)	Speaker	Syntactic rule	Yield
185.105	follower	VP → VBG PP	*keeping on the edge of the page*
185.363	follower	PP → IN NP	*on the edge of the page*
185.490	follower	NP → AT NN	*the edge*
185.490	follower	NP → NP PP	*the edge of the page*
185.692	follower	PP → IN NP	*of the page*
185.729	follower	NP → AT NN	*the page*

and target (DIST) in terms of utterances, while the following experiments model it in seconds.[2] Cumulative priming by a stimulus that is repeated several times is not captured by this statistical model.

We distinguish *comprehension-production* (CP) priming, where the speaker first comprehends the prime (uttered by his/her interlocutor) and then produces the target, and *production-production* (PP) priming, where both the prime and the target are produced by the same speaker. This distinction is encoded in the factor CP, which is coded as 1 for between-speaker CP priming, and 0 (base case) for within-speaker PP priming.

A predictor ln(FREQ) is included to control for the frequency of the repeated syntactic rule in the corpus, as the log-transformed rule frequency normalized by corpus size. Frequency is an important covariate in many psycholinguistic models and has long been suspected to interact with priming (e.g., Scheepers, 2003).

In summary, our model demonstrates a priming effect by observing a decay, that is, a negative parameter for ln(DIST). How strong this decay is gives us an indication of how much repetition probability we see shortly after the stimulus (prime) compared to the probability of chance repetition—without ever explicitly calculating such a prior. We define the strength of priming as the decay rate of repetition probability, from shortly after the prime to 15 seconds or 25 utterances afterward (predictor: ln(DIST)). Thus, we take several samples at varying distances (d), looking at cases of structural repetition, and cases where structure has not been repeated.

Experiment 1: Repetition in Corpora

While controlled experiments have shown syntactic priming, we first aim to demonstrate a sensitive method that can quantify and contrast priming magnitudes in corpora. We will examine two types of text: (a) spontaneous conversation, that is, in a situation where the semantics of the dialogue are not controlled, and (b) task-oriented dialogue, where interlocutors collaborate to achieve a common goal.

Method

We use two datasets in this experiment and build two separate statistical models. Short-term priming effects are measured as described previously. The first dataset is *Switchboard* (Marcus et al., 1994), a corpus of spontaneous spoken telephone dialogues among randomly paired, North American English speakers who were given a general topic to discuss, but were otherwise unrestricted. The corpus contains 80,000 transcribed utterances were annotated with phrase structure trees (Marcus, Kim, Marcinkiewicz, MacIntyre, Bies, Ferguson, Katz and Schasberger, 1994), yielding 472,000 phrase structure rules with 4,700 distinct rules. Words in this portion of the corpus, included in the Penn Treebank, were time-tagged (Carletta et al., 2004). After extracting all potential repetition cases, the data were balanced by re-sampling, yielding an equal number of repetition and non-repetition cases.

The second dataset is the *HCRC Map Task* corpus (Anderson et al., 1991), which consists of 128 task-oriented dialogues containing 20,400 utterances, using 759 different phrase structure rules. Using exactly the same methodology as for Switchboard, we extracted 157,000 rules.

Results

Two regression models were fitted, one to each dataset (Table 2). They contain the ln(DIST) covariate to estimate priming levels (negative effects indicate stronger priming), ln(FREQ) for the effects of frequency, and a factor CP (to identify comprehension-production priming between speakers).

In *Map Task*, ln(DIST) reliably predicts declining rule repetition ($\beta = -0.073$, $p < 0.0001$). Repetition of a rule becomes less likely as the distance measured in utterances from the first occurrence increases: ln(FREQ) interacts reliably with ln(DIST) ($\beta = 0.043$, $p < 0.0001$). In *Switchboard*, ln(DIST) also predicts declining rule repetition ($\beta = -0.080, p < 0.0001$), and the effect is reduced by increasing frequency. Prime Type CP (priming between speakers) does not interact with the decay coefficient for ln(DIST).[3] ln(FREQ) interacts with ln(DIST) ($\beta = 0.057$, $p < 0.0001$),

[2] We aim to show broad applicability of the method, but see time as the most reliable and neutral basis for decay. Reitter (2008) contains further experiments varying this metric.

[3] The resulting estimate for ln(DIST) in our model (for a syntactic rule of average frequency) would be −0.080 for PP (odds ratio: 0.92), but −0.080 − 0.017 (odds ratio 0.91) for CP priming. Because a negative β indicates decay, this indicates CP and PP priming in Switchboard.

Table 2
Two regression models of short-term rule repetition (Experiment 1). Prime-target distance in utterances. All continuous predictors were centered; CP was coded as 1, PP is the base case. Response variable (repetition probability), effect sizes (β) and standard errors (SE) in logits. Random effects of intercept and slope (distance), grouped by utterance. Maximum accepted correlation between covariates 0.2; CP was residualized. p-values (according to $|z|$), < 0.05 *, < 0.0001 ***.

	MapTask				Switchboard			
Covariate	β	OR	SE		β	OR	SE	
Intercept	−1.721	0.18	0.011	***	−1.079	0.34	0.025	***
ln(Dist)	−0.073	0.93	0.011	***	−0.080	0.92	0.012	***
ln(Freq)	0.722	2.06	0.01	***	0.884	2.42	0.006	***
CP	−0.684	0.50	0.013	***	−0.176	0.84	0.011	***
ln(Dist):CP	−0.018	0.98	0.019		−0.017	0.98	0.014	
ln(Dist):ln(Freq)	0.043	1.04	0.011	***	0.057	1.06	0.006	***

which suggests that repetition probability decreases less quickly for rules with high frequencies. That is, we find less priming for more common rules.

Discussion

A speaker is more likely to use a syntactic rule shortly after using the same rule. The closer prime and target are to one another, the stronger the preference is to repeat. Priming occurs both within a speaker (PP) and between speakers (CP), and it decays rapidly. The method to quantify priming by estimating the decay effect was developed initially for the Switchboard corpus; Map Task was not used to design or tune the regression modeling methods.

The priming effect obtained in these corpora confirms experimental results by Bock and Griffin (2000) and Branigan et al. (1999). These studies find syntactic priming over short and longer time periods.[4] The decay we observe is remarkable: repetition rates reach levels indistinguishable from the prior after about 5-6 seconds. At first glance, this contrasts with Szmrecsanyi's (2006, p. 188) results, who finds that future marker choices (*will* vs. *going to*) decay only after 140 words (which would be approximately 45 seconds at a speech rate of 180 words/min). However, as Szmrecsanyi points out, due to the logarithmic nature of the forgetting function, most of the priming effect "declines within an interval of 10 words (...), equivalent to ca. 5 seconds of speech." With our data, a log-linear model (for distance) yielded a better fit than a linear-linear one[5], which is compatible with general models of memory (Anderson et al., 1998).

The models produced for Switchboard and Map Task cannot be used to quantify the strengths of syntactic priming; they just show the decay effects separately for the two corpora. In the next experiment, we compare priming between the corpora.

Experiment 2: Priming and Decay Over Time in Different Genres

In this section, we develop the first of two hypotheses designed to test the IAM or some of its assumptions.

The IAM suggests that priming benefits speakers in conversation. At the same time, we observe that independently fitted statistical models appear to paint a different picture of priming in spontaneous conversation, as opposed to priming in task-oriented dialogue.

The test of the IAM we put forward presupposes *rationality* in cognitive processes, that is, that variation in an individual's linguistic processes tends to optimize the communicative or situational outcome. If we accept this as a general principle (Anderson and Milson, 1989; Chater and Oaksford, 1999), then the IAM predicts that if speaker's priming levels vary at all with dialogue purpose, they tend to vary such that task-oriented dialogue shows stronger priming than less goal-driven interaction, i.e., spontaneous conversation or small talk.

Let us briefly consider the alternatives. First, if priming is the result of a mechanistic memory effect that is not influenced by dialogue purpose or contextual working memory contents, then we should not observe any difference in priming between the dialogue genres. Second, if we do find different priming levels, and we see more priming in spontaneous conversation, we would interpret this as a violation of the IAM prediction or even rationality as a whole.

The differences in dialogue situation may have affected priming levels through a different mechanism than IAM. Speakers may have tailored their utterances to match the needs of their audience: In the experimental design that led to the Map Task data, participants were in the same room and half of the pairs could make eye contact. From an *audience design* perspective, the richer communication channel may have led them to reduce their levels of adaptation in Map Task. This is contrary to what would be expected under the IAM.

Next, we describe the Map Task in detail. This corpus will be used throughout the remainder of this paper.

[4] The effect of CP on bias may be related to general levels of speaker idiosyncrasies, i.e., increased chance repetition within speakers. Fitting the main effect controls for that.

[5] Applying the Akaike Information Criterion, the model in Table 3 would be exceedingly unlikely, if it employed linear distance instead of log-linear distance ($p < 0.0000$).

The Map Task. Like Switchboard, the Map Task is a corpus of spoken, two-person dialogues in English. Unlike Switchboard, the Map Task dialogues are *task-oriented dialogues*, in which interlocutors work together to perform a task as quickly and efficiently as possible. In each trial, the two speakers sat opposite one another and each had a map, which the other could not see. One of them, the *instruction giver*, had a map with a route drawn on it; the other participant, the *instruction follower*, had no route drawn on her map. The speakers were told that their goal was to reproduce the Instruction giver's route on the Instruction follower's map. The maps were not identical, and before they began the task the participants were told explicitly that their maps may differ in some respects, and that they could say whatever was necessary to complete the task. It was up to the participants to discover how the two maps differed (see Figures 4 and 5).

All maps consisted of landmarks represented as line drawings which are labelled with their intended name. All map routes began with a starting point, which was marked on both maps, and an end point, which was marked only on the giver's map. Landmarks along the map alternated between those that appeared on both maps and those that appeared on only one map. For each map, 8 landmarks appeared on both maps, 4 on only the giver's map, and 3 on only the follower's map. In addition, some landmarks (typically one per map pair) had different names on the two maps. These names were identical in form and location but had different labels on the two maps (e.g., *mill wheel* vs. *old mill*). Finally, 2 landmarks appeared twice on the giver's map, once in a position close to the route and once in a position more distant from the route. The follower had only one repeated landmark, which was distant.

Each subject participated in four dialogues, twice as instruction giver and twice as instruction follower. The spoken interactions were recorded, transcribed and syntactically annotated with phrase structure grammar.[6]

Method

We pool the two datasets (*Switchboard* and *Map Task*), distinguishing them via a factor SOURCE. The methodology to quantify priming levels is the same as for the previous experiments, except that the DIST covariate is now measured in seconds instead of utterances (the notion of utterance is not the same in each corpus, and average utterance length differs).[7]

Results

Refer to Table 3. The estimate for ln(DIST) describes the slope of repetition probability over time for the baseline condition, that is, in Switchboard. We find a main effect of ln(DIST) ($\beta = -0.165$, $p < 0.0001$). This indicates priming in Switchboard. ln(DIST) interacts with MAPTASK ($\beta = -0.058$, $p < 0.001$), indicating reliably stronger priming in Map Task. As before, ln(DIST) also interacted with ln(FREQ) ($\beta = 0.092$, $p < 0.0001$), i.e., priming is stronger for less frequent rules. An interaction between ln(DIST), CP and MAPTASK ($\beta = -0.106$, $p < 0.005$) documents that there is a larger gap between CP and PP priming in Map Task than in Switchboard. (Between-speaker priming is strong in task-oriented dialogue, but not in spontaneous conversation, first reported in Reitter et al. (2006)).

Figure 2 contrasts different effect sizes, that is, estimates of priming strengths (ln(DIST) interactions) for the four factor combinations of CP and SOURCE. The post-hoc confidence intervals as well as the model suggest that priming between speakers (comprehension-production) may be stronger than priming within a speaker (production-production) for Map Task only.

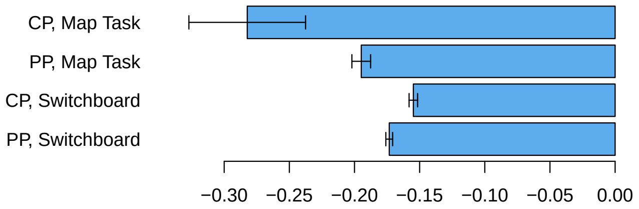


Figure 2
Relative Decay effect sizes in logits for ln(DIST) with different combinations of CP and SOURCE factors and average residual frequency, based on model shown in Table 3. Longer bars indicate stronger decay and priming. Error bars show standard errors.

To illustrate the relative magnitude of the effects, we give conditional repetition probabilities in our data. Recall that these data were resampled to provide an overall higher proportion of repetitions to facilitate model fitting. The average repetition probability is 0.170 for all samples from Map Task, and 0.156 in Switchboard. In Map Task, in the first two seconds after a prime, over all syntactic rules, and between speakers, repetition probabilities are $p = 0.219$, and at distances of 8–10 seconds, 0.143. For Switchboard, they are 0.165, and 0.141, respectively. That is, repetition is not only more common in Map Task, but, crucially, its drop-off is greater.

The model itself can make predictions. To derive these, effects and interactions have to be combined in logit-space, taking into account centering and log-transformations. (We assume average random effects.) For illustration purposes, we chose the rule $S \rightarrow PP\ S$, which licenses a clause beginning with a preposition (such as *below that bend there is an abandoned cottage*). The frequency of this rule is 63% of that of the mean frequency (which is still more common than 87% of all rules in Map Task). In the Map Task data, for priming between speakers, the model's prediction for repetition of that rule in regular probability space is $p = 0.173$ at a prime-target distance of one second, and 0.120 at nine seconds. For Switchboard, these values would be 0.158 and 0.109, respectively. (Many constructions most commonly examined in

[6] Many other types of annotation are also available. See http://www.hcrc.ed.ac.uk/maptask/ for a description and instructions of how to obtain the corpus.

[7] Elsewhere, we have documented that time-based vs. utterance-based analysis does not confound the comparsions between the corpora Reitter (2008).

priming studies, such as passives, are very rare in our speech corpora.)

Discussion

The model based on temporal distance confirms the earlier model based on utterances. The basic result from Experiment 1 holds: there is syntactic priming in both corpora.

The experiment lends support to our hypothesis: evidence for stronger syntactic priming when speakers engage in purpose-driven conversation. Priming is stronger in the task-oriented Map Task corpus than in the spontaenous conversations of Switchboard.

Discussion: Results and Methods

In summary, reliable syntactic priming effects can be detected in natural dialogue for general syntactic rules instead of selected constructions. We model syntactic priming as the decay of repetition probability of syntactic rules, both in the course of linguistic activity (utterances), and over time.

Both of the corpora of spoken dialogue that we investigated showed an effect of distance between prime and target on syntactic repetition probability, thus providing evidence for a structural priming effect for arbitrary syntactic rules. In both corpora, we also found reliable effects of both production-production (PP) priming (self-priming) and comprehension-production (CP) priming. With the clear PP priming effect in spontaneous conversation, we also add a new finding compared to Dubey et al. (2005), who did not detect reliable evidence of adaptation within speakers in Switchboard for selected syntactic rules in coordinate structures.

In the Map Task corpus, which consists of task-oriented dialogues, we find evidence for stronger overall priming than in Switchboard, a corpus of spontaneous conversation. We consider this effect supporting evidence for the Interactive Alignment Model (Pickering and Garrod, 2004). According to the IAM, what we observe is the reciprocal boosting of syntactic priming and the alignment of the situation models present in task-oriented dialogue. The interaction partners synchronize their situation models in the task-oriented setting, which co-occurs with cross-speaker priming (CP) on other communicative levels. CP priming appears to be enhanced by the need for a shared situation model. Recurring coordination moves enable speakers to make fine-grained distinctions of the path described, and these may provide an explanation for increased local repetition. As a concurrent explanation, semantic and lexical material that occurs in clusters may also have facilitated local syntactic repetition.

We concede that dialogues in the two corpora differ greatly with respect to the overall goals of the speakers, their mode of interaction, the durations of their turns, their language registers and their linguistic variability. While the underlying, decay-based methodology can be expected to be robust with respect to general differences in language, it is still unclear which differences between the corpora actually caused priming to be stronger in Map Task. The next experiments address this concern. We will examine only data from the Map Task corpus, which was collected under well-controlled conditions. We also broaden our view to distinguish short-term and long-term adaptation, and to evaluate to what extent task success can be predicted and estimated based on lexical and syntactic adaptation.

Large corpora present us with an opportunity to evaluate small effects and multiple interactions. Yet, data points gained from linguistic corpora are never independent (Kilgarriff, 2005). For instance, a single utterance will typically yield multiple syntactic data points, but of course, the choices of syntactic constructions in a sentence depend heavily on one another. In the corpus study presented here, care is taken to group such linguistic interdependencies in the (random effects) models. A further issue arises due to sub-languages resulting from corpus choice, genre, or speaker. The model structure controls for such variation by contrasting primed and non-primed samples within the same corpus, and by using decay as the target metric to measure priming.

A final methodological concern is coherence: adjacent utterances do not jump from topic to topic—instead, they form clusters or discourse segments that are topically coherent (Grosz and Sidner, 1986). Clustering may be present as a result of convention or processing constraints, but it may also be introduced by the task as it is in Map Task, where the path is typically drawn step-by-step, with the area around one landmark being discussed at a time. Could clusters be responsible for the short-term priming effect, producing more repetition inside a cluster than outside (and further away)? This potential confound would affect the short-term priming, but not the long-term adaptation measure. Most importantly, topic chains are reflected primarily in lexical choice, and only indirectly (e.g., via topic status) in syntactic configuration.

Experiment 3: Task Success and Short-Term Priming

Under the IAM, we expect successful dialogues to show more priming than unsuccessful ones. To test the IAM hypothesis, we assume that success at the Map Task is an indicator of aligned situation models. The next experiment is designed to detect co-variance of short-term priming and task success.

Method

The Map Task consists of re-tracing a defined route according to the interactive description provided by the other interlocutor. So, task performance is measured in terms of how far the route that the follower has drawn deviates from the route shown on the giver's map. To compute this for each dialogue, the developers of the Map Task corpus overlaid the giver's map on the follower's map and computed the area covered in between the paths (PATHDEV). Task success is then defined as the inverse of PATHDEV.

We correlate short-term priming levels in each dialogue with path deviation. The underlying model is the same

Table 3
The regression model for the joint dataset of Switchboard and Map Task (Experiment 2), prime-target distance Dist in seconds. This is the minimal model without unjustified covariates. All variables were centered. Random variables for intercept and ln(Dist), grouped by utterances. All continuous variables were centered; CP and ln(Freq) are residuals after regressing out effect of ln(Dist); resulting coding: MapTask: 0.51 vs. base case (Switchboard) −0.49, CP: 0.65 vs. base case (PP): −0.35. Fixed-effect correlations between all variables was lower than 0.25. ANOVA F-values shown.

Covariate	β	OR	SE	F	z	$p(>\lvert z\rvert)$
Intercept	−2.096	0.12	0.010		-200.6	< 0.0001 ***
ln(Dist)	−0.195	0.84	0.011	82.7	-17.3	< 0.0001 ***
CP	−0.263	0.83	0.014	304.5	-18.8	< 0.0001 ***
MapTask	0.054	1.06	0.015	2.6	3.49	< 0.001 **
ln(Freq)	0.759	2.14	0.007	10388.8	102.2	< 0.0001 ***
ln(Dist): CP	−0.033	1.02	0.019	5.4	-1.77	< 0.10
ln(Dist): MapTask	−0.058	0.94	0.017	12.6	-3.35	< 0.001 **
CP: MapTask	−0.166	0.85	0.028	35.1	-5.93	< 0.0001 ***
ln(Dist): ln(Freq)	0.092	1.10	0.009	113.3	10.65	< 0.0001 ***
ln(Dist):CP:MapTask	−0.106	0.90	0.037	8.2	-2.87	< 0.005 **

as in Experiment 1, except that an interaction of Dist and PathDev is included to measure this relationship. Prime-target distance ln(Dist) is measured in time (seconds). Under the IAM, we expect there to be more priming with greater task success. As Dist is lower for stronger priming, and PathDev is lower for more successful dialogue outcomes, we expect a positive estimate for this interaction.

Results

Table 4 shows the full model. As before, short-term priming is reliably correlated (negatively) with ln(Dist), hence we see a decay and priming effect (ln(Dist), $\beta = -0.150$, $p < 0.0001$). Notably, however, path deviation and short-term priming did not correlate. We tested for reliable PathDev and ln(Dist) interactions, separately for PP and CP situations via contrasts. In neither case did we find a reliable interaction.

Discussion

We have shown that although there is a clear priming effect in the short term, the size of this priming effect does not correlate with task success. But does this indicate that there is no strong functional component to priming in the dialogue context? There may still be an influence of cognitive load due to speakers working on the task, or an overall disposition for higher priming in task-oriented dialogue: Experiment 2 points to stronger priming in such situations. Our results are difficult to reconcile with the model suggested by Pickering and Garrod (2004), if we take short-term priming as the driving force behind the IAM.

A hypothetical explanation of our failure to find the priming–task success correlation is that short-term priming decays within a few seconds. It is questionable to what extent such a brief effect helps interlocutors align their situation models. In the Map Task experiments, one of the linguistic devices where *lexical* alignment is expected to make a difference is reference to landmarks. Do interlocutors need to refer to landmarks every few seconds? Syntactic priming forms part of alignment of such references through the internal structure of noun phrases that identify the landmarks. Syntactic devices may also be avoided within the early period of rapid decay of repetition probability that we observe. We hypothesized that the syntactically more complex descriptions of how to circumnavigate the landmarks would be repeated on the order of several times a minute, but not commonly within 5–10 seconds. An analysis of the dialogues, however, showed that reference is used much more frequently than we expected. The task lends itself to a clustering of references to the same landmark, as speakers describe the route step by step. Thus, our hypothetical explanation cannot be corroborated.

An alternative explanation comes from the empirical literature: there are two distinguishable, but interacting adaptation effects. A fast, short-term priming effect, and long-term adaptation that persists (Ferreira and Bock, 2006). In the cognitive model we proposed in Reitter et al. (2011), short-term priming is enhanced by semantic material held in short-term memory, but memories of syntactic structures are reinforced and become increasingly more accessible with each use. This provides an explanation for the observed stronger priming in task-oriented dialogue. In the next experiment, we seek to link task success to long-term adaptation.

Experiment 4: Task Success and Long-Term Adaptation

Interactive alignment is a process that happens on the time-scale of minutes: speakers establish a common reference system in the long run. This process may not as initially thought be based on short-term priming. Pickering and Garrod (2004) do not detail the longevity of the priming ef-

Table 4
The full regression model for the Map Task dataset (Experiment 3). CP indicates between-speaker (comprehension-production) priming; PP is within-speaker priming. The scale of PathDev is in mm^2 to indicate the area of path deviation in the Map Task; as centered, it ranges from −64 to +159.All covariates were centered; fixed-effect correlations between all centered variables was lower than 0.2. Model ANOVA corroborate the significance of parameter tests (F-values shown).

Covariate	β	SE	F	z	$p(>\lvert z \rvert)$	
Intercept	−1.747	0.174	0.014		-127	< 0.0001 ***
ln(Dist)	−0.150	0.860	0.014	86.7	-10.5	< 0.0001 ***
CP	−0.364	0.695	0.020	277.6	-18.2	< 0.0001 ***
PathDev	0.0002	1.000	0.0002	0.153	0.81	0.42
ln(Freq)	0.700	2.013	0.012	3557	59.9	< 0.0001 ***
ln(Dist):CP	0.911	−0.093	0.024	14.5	-3.91	< 0.0001 ***
ln(Dist):ln(Freq)	0.080	1.083	0.013	39.4	6.27	< 0.0001 ***
ln(Dist):PathDev/PP	0.000	0.0000	0.0003	0.03	0.07	0.95
ln(Dist):PathDev/CP	0.000	0.0001	0.0004		-0.21	0.84

fects supporting alignment. Is is unclear whether alignment is due to the automatic, classical *priming* effect, or whether it is based on a long-term effect that is possibly related to implicit learning (Bock and Griffin, 2000; Chang et al., 2006; Kaschak et al., 2011a). The next experiment investigates the latter possibility. Analogous to the previous experiment, we hypothesize that more long-term adaptation relates to more task success.

Method

For structural priming[8], two repetition effects have been identified. Classical structural priming effects are strong: around 10% for syntactic rules (Reitter et al., 2006). However, they decay quickly (Branigan et al., 1999) and reach a low plateau after a few seconds, which makes the effect seem similar to semantic priming. What complicates matters is that there is also a different, long-term syntactic *adaptation* effect that is also commonly called (repetition) priming.

Structural adaptation has been shown to last longer, from minutes (Bock and Griffin, 2000) to several days. Lexical boost interactions, where the lexical repetition of material within the repeated structure strengthens structural priming, have been observed for short-term priming, but not for long-term priming trials where material intervened between prime and target utterances. Thus, short- and long-term structural adaptation effects may well be due to separate cognitive processes, as argued by Ferreira and Bock (2006).

After the initial few seconds, structural repetition shows little decay, but can be demonstrated even minutes or longer after the stimulus. To measure this type of adaptation, this method looks at repetition of syntactic rules over whole document halves, independently of decay.

This method splits each dialogue in half. Analogous to the short-term priming model, we define repetition as the occurrence of a prime within the first document half (Prime), and sample rule instances from the second document half. To rule out short-term priming effects, 10-second portion in the middle of the dialogues is excluded.

In order to distinguish adaptation from overall, random repetition of syntactic rules, we contrast dialogue halves stemming from single dialogues with dialogues halves taken from two different dialogues. A factor SameDoc distinguishes between the two cases. For SameDoc=0, we combine dialogue halves stemming from different dialogues[9]; for SameDoc=1, the dialogue halves stem from the same dialogue. Thus, our model estimates the influence of preceding context on rule repetition. The goal is now to establish an effect of SameDoc on repetition.

Using the same data as in Experiment 3, task success is inverse path deviation PathDev as before, which should, under IAM assumptions, interact with the effect estimated for SameDoc. The response variable is Prime, indicating whether a rule is repeated.

Results

As seen in Table 5, SameDoc showed a reliable, positive effect ($\beta = 3.303$, $p < 0.0001$), which means we see long-term adaptation. This generalizes previous experimental priming results in long-term priming. The effect interacted reliably with the path deviation scores (SameDoc:PathDev, $\beta = -0.624$, $p < 0.05$). Thus, we find a reliable correlation of task success and syntactic priming. Greater path deviations relate to weaker priming.

[8] In both production and comprehension, which we do not distinguish further for space reasons.

[9] This is a control condition; particularly if applied to lexical repetition, topicality can lead to repetition that is higher than would be sampled from large corpora in the same language (see Church (2000), which inspired the methodology used here)

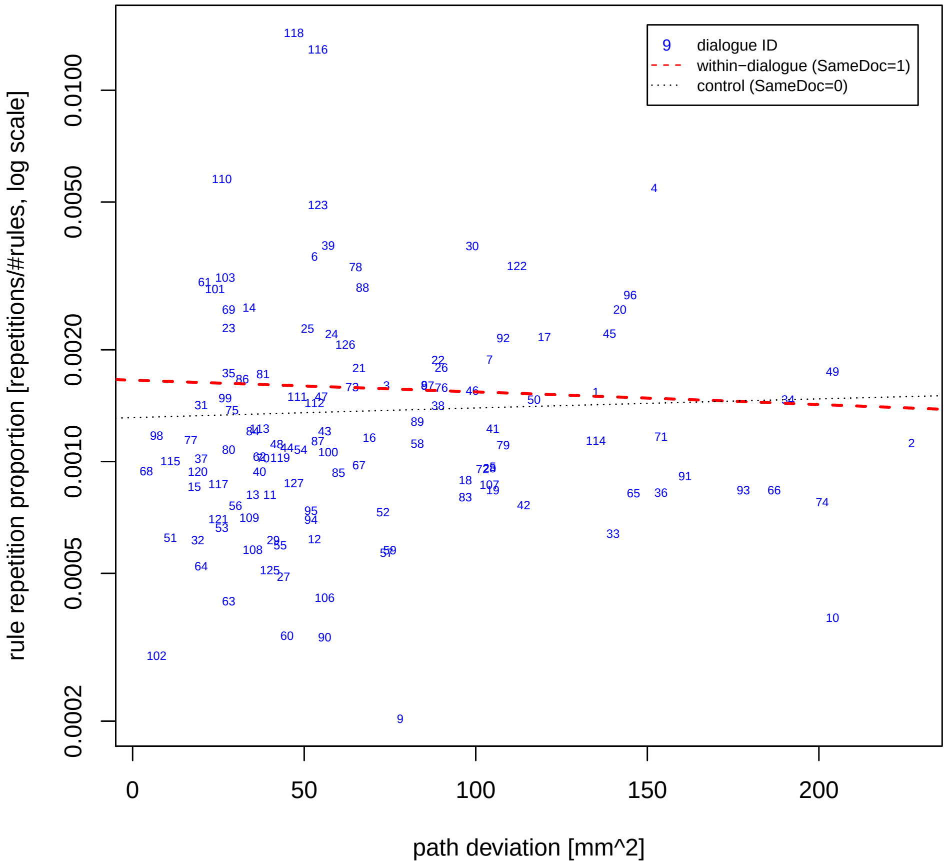


Figure 3
Alignment vs. task success for each dialogue in Map Task.

Covariate	β	OR	SE	F	z	$p(>\|z\|)$
Intercept	2.722	15.2	0.036		75.5	< 0.0001 ***
ln(FREQ)	1.499	4.48	0.016	478	13838	< 0.0001 ***
SAMEDOC	1.064	2.90	0.048	478	22.0	< 0.0001 ***
PATHDEV	−0.001	1.00	0.001	2.27	-1.03	$= 0.3$ ***
ln(FREQ):SAMEDOC	−0.001	0.9990	0.0002	16.5	-4.37	< 0.0001 ***
SAMEDOC:PATHDEV	−0.002	0.9977	0.001	6.32	-2.51	< 0.05 *

Table 5
The logistic regression model for the Map Task dataset (Experiment 4). The scale of PATHDEV is in mm^2 to indicate the area of path deviation in the Map Task; as centered, it ranges from −64 to +159. Thus, β and odds ratio (OR) for the critical parameter apply to a single mm^2 in difference. All covariates were centered; fixed-effect correlations between all centered variables was lower than 0.2. Model ANOVA corroborate the significance of parameter tests (F-values shown).

The normalized rule frequency ln(FREQ) did not interact with SAMEDOC ($\beta = -0.044$, $p = 0.35$). Such an interaction also could not be found in a reduced model with only SAMEDOC and ln(FREQ). The interaction was removed from the model.

The effect of long-term adaptation can be visualized in a simple way. In Figure 3, the proportions of repeated to novel syntactic rules in each dialogue are related to path deviation, contrasting within-dialogue and between-dialogue repetition (control).

Discussion

Speaker pairs' long-term syntactic adaptation is correlated with the synchronization of their routes on the maps. This is exactly what one would expect under the assumption of the IAM. We find no evidence for stronger long-term adaptation of rare rules, which may point out a qualitative difference to short-term priming. Taken without theoretical motivation, the results do not imply causality. However, task success is unlikely to cause increased priming, as participants in Map Task were not told whether they were on the "right track". Mistakes, such as passing a landmark on its East and not on its West side, were made and went unnoticed. The repetition effect that contributes to prediction accuracy is long-term syntactic adaptation as opposed to short-term priming.

Predicting Task Success

So far, we have put forward a case for a link between syntactic adaptation and task success. However, the IAM spans more than the syntactic level. Lexical priming is also part of the priming cascade. In the following, we establish the predictiveness of linguistic similarity for task success with a more complex model that includes lexical features. Second, we demonstrate the computational applicability of our findings. In an application, an automatic estimate of task success could help evaluate conversations among humans. In human-computer dialogues, predicting the task success after just a first few turns of the conversation could avoid disappointment with the system by switching dialogue strategies or by passing poorly performing automated calls on to a human operator.

Experiment 5: The success prediction task

In this section, we define a general task that predicts conversational success from textual features. The task we set for ourselves requires that *success is estimated* from the contents of an entire dialogue. All linguistic and non-linguistic information available may be used. This task reflects post-hoc analysis applications, where dialogues must be evaluated without an independent success measure being available for each dialogue. This covers cases where, for example, it is unclear whether a call center agent or an automated system actually responded to the call satisfactorily. In the next section, we describe a statistical approach that uses repetition effects to implement this task.

Method

We use a standard machine-learning algorithm, a Support Vector Machine (SVM), which acquires a model from data that, in our case, predicts task success from a set of features. It can do so for a range of data points, each of which consists of a task success value (as given for the dialogue) and a set of features. The SVM uses features representing lexical and syntactic repetition information. We include data points as snapshots at each 10-second interval in each dialogue, with features encoding the cumulative lexical (LEXREP), syntactic (SYNREP) and character-based (CHARREP) repetition counts up to that point in time. We do not distinguish repetition between and within speakers. Syntactic repetitions are based on phrase-structure rules as before; lexical repetitions are based on words. A time stamp and the total numbers of constituents and characters are also included (LENGTH). This way, the model can work with repetition proportions rather than the absolute counts. Unlike in the previous, hypothesis-driven experiments, the emphasis here is on task performance rather than on the model's parsimony or on estimates that can be interpreted with respect to the initial hypothesis. The SVM is trained for regression with a radial basis function kernel ($\gamma = 5$), using the PATHDEV score as output.

Evaluation

A suitable evaluation measure, the classical R^2, indicates the proportion of the variance in the actual task success score that can be predicted by the model. All results reported here are produced from 10-fold cross-validation, using 90% training / 10% test splits of the dialogues. No full dialogue was included in both test and training sets.

The results (Table 6) indicate that ALL repetition features together with the LENGTH of the conversation, account for 17% of the total score variance. The repetition features improve on the performance achieved from dialogue length alone (9%). When the syntactic repetition feature is taken out, we achieve 15% in explained variance. The baseline was implemented as a model that always estimates the mean score. It should, theoretically, be close to 0.

Discussion

Linguistic repetition serves as a good predictor of how well interlocutors will complete their joint task. The features used are relatively simple: provided there is some syntactic annotation, as available from an automatic syntactic parser, rule repetition can easily be detected. Even without syntactic information, lexical repetition already goes a long way. In applications where no syntactic annotation is provided, part-of-speech tag n-grams (which are easy to obtain reliably) show the same decay-based priming effects (Reitter and Keller, 2007). Obviously, linguistic information alone does not explain the majority of the task-solving abilities. Communicative strategies should play a role, as does the dyad's understanding about how much precision is required. Some subject pairs may be more or less motivated to do well (HCRC Map Task participants were not incentivized). Despite the noise introduced by such factors, we do find consistently that repetition and a tendency to adapt are associated with task success.

These application-oriented results strengthen our initial hypothesis of the link between the tendency to repeat choices in language production and the success of the communicative process as a whole. Choices are no longer limited to sentence structure. If one accepts the sometimes explicit rather than implicit lexical choices as data points for adaptation, then the SVM model can lend support to the IAM at these other representational levels. The results are compatible with a view

Table 6
Portion of variance explained (R^2). ALL includes the features SYNREP, LEXREP, CHARREP.

ALL and LENGTH	**0.17**
ALL without SYNREP	0.15
ALL without LEXREP,CHARREP	0.09
LENGTH ONLY	0.09
Baseline	0.01

that sees a predisposition in speakers to adapt to one another more or less (Gill et al., 2004), and, as in the IAM, that positive adaptation ultimately leads to task success. Such a correlation between syntactic adaptation and task success is visible early on in the dialogues, more so than any correlation of lexical adaptation and task success.

General Discussion

Given the correlation spanning the IAM hierarchy of priming from syntactic choice to task success, we can revisit the initial experiments (1–2), where we develop a methodology to measure short-term priming in corpora. We find a reliable difference in syntactic priming between task-oriented dialogue and spontaneous conversation in two distinct corpora. These experiments tested a prediction; they are not post-hoc analyses. Notwithstanding, the corpus-based design is non-manipulative and cannot determine the exact cause of the difference in priming. If short-term priming does not influence task success, why would there be more short-term priming in task-oriented dialogue than in spontaneous conversation? Could participants have explicitly controlled their alignment? We chose syntactic priming as an indicator precisely because syntactic choices are usually implicit. We also chose to use naturalistic dialogue in corpus data rather than laboratory studies to avoid the possibility that syntactic decisions would become evident to participants. During the data collection experiments that led to the Switchboard and Map Task corpora, participants saw no contrastive use of syntactic alternations in any experimental material that may have led them to make explicit decisions about the structure of their speech.

Both measures, of short-term and long-term adaptation, control for baseline levels of repetition. The short-term priming measure is based on decay rather than repetition, for this reason, and the long-term adaptation measure compares repetition after possible adaptation to repetition after adaptation was impossible. Dialogue partners were matched by the experimenter (rather than self-selected), and unlike other priming studies, we take a broad variety of syntactic structures into account.

Could it be that semantic activity in task-oriented dialogue facilitated cue-based memory retrieval (priming)? The last of these explanations depends on the nature of semantic processing that we expect to find in task-oriented dialogue. In the Map Task experiments, listeners actively processed what was being said, keeping a subset of a small set of items such as landmarks in working memory, because the task demanded just that. In the conversations recorded in the Switchboard corpus, interlocutors were not required to remember or process much of the content discussed, or when they do, more varied content resides for a briefer period of time in working memory. In Reitter et al. (2011), we propose a mechanism for short-term priming that depends on spreading activation of lexical (and thus also semantic) material. Indeed, we suggest that more intense semantic processing leads to more lexical material being retained in working memory, serving as cues in the retrieval of associated syntactic structures. This is what may have caused strong priming in task-oriented dialogue, and presumably quite generally in "engaged" dialogue. That said, this is only one of several possible accounts of the mechanisms of syntactic adaptation; a variety of mechanisms would be compatible with the functional claims we make (cf., Healey (2011)).

The link between syntactic adaptation and task success is corroborated by a recent study on lexical alignment. In a task given to dyads (Fusaroli et al., 2012), participant pairs who aligned in their word choices also did better in the given task. This effect, however, was seen only for task-relevant vocabulary, and not for general lexical alignment.

The fact that short-term priming and long-term adaptation differ qualitatively is relevant from an architectural viewpoint. It suggests that there is more than one cognitive basis for these repetition effects: if there was only one, we would expect short-term priming and long-term adaptation to covary with variables such as task success (Ferreira and Bock, 2006; Reitter et al., 2011). Whereas short-term priming appears to be modulated by cognitive processes reflecting the dialogue goals (i.e., genre), it is not the short-term syntactic priming mechanism that leads to high-level alignment, as Experiments 5 and 6 demonstrate. Alignment that aids interlocutors in performing their joint task is not the result of short-term priming or cue-based memory retrieval. The search for alternative mechanisms behind the link between linguistic adaptation and task success has pointed out a number of dialogue metrics, some of which are known to correlate with task success. The Map Task, with its divergent maps makes it beneficial for subjects to seek confirmation of their location and the surrounding landmarks before giving instructions about how to circumnavigate them: Anderson and Boyle (1993) found the number of yes/no questions to be positively correlated with task success (measured as negative path deviation). Other means of interaction, such as the use of non-verbal communication, led to more efficient dialogue (Boyle et al., 1994). This was a controlled variable in the Map Task corpus collection experiment, as was the familiarity among subject pairs. Could such metrics and variables predictive of success hold clues for an alternative explanation of the alignment effect? Table 7 shows the absence of correlation of a number of such variables known or suspected to be predictive of task success and a basic repetition metric. The HCRC Map Task experiment controlled these variables (eye-contact, familiarity between speakers). To correlate the variables with repetition, we collected counts of rule repetition between first and second dialogue halves and normalized them by the number of overall rules and the expected effect of their frequencies, as a simple, exploratory measure of long-term adaptation. We found no evidence for any latent mechanisms that would explain or confound alignment. (Even the presence of yes/no questions is unlikely to lead to increased syntactic repetition.) Because the measures are high-level statistics, we examined a sample of the conversations, looking for intentions that could explain the repetitive use of lexical or syntactic forms.

Simple repetition may communicate agreement with, or respect for, the interlocutor. Indeed, the rate of adaptation may sometimes depend on the speaker's assumptions about the recipient of the message. For example speakers adapt their lexical and syntactic choices more to a (presumed) inferior computer interlocutor, less so with a (presumed) advanced computer, and least with a human (Branigan et al., 2010). It is reasonable to assume an implicit control mechanism.

A competing explanation for alignment in specific situations is that repetition conveys meaning. In particular, we might interpret repetition as a culturally established normative convention. One example of repetitive use of lexical and syntactic material was given in the introduction (p. 2): in this excerpt, the instruction giver repeats the follower's earlier phrase "like a tick shape" with the parallel construction "like an oval shape". However, the last phrase the follower used was "the shape of a tick", and this was the syntactic form the giver attempted to use first. It is possible that this first use was adapted to the most active, available syntactic form ("the shape of a N"). Consider a similar example from the MapTask corpus:

g: *that's how you go just just* **go round the top**
f: you actually **go round the side** *where the where the wheel is*

Such syntactic repetition can mark semantically contrastive use. As such, repetition is a linguistic device indicative of pragmatic intentions, as is non-repetition in situations where repetition would be expected. Our methodology does not distinguish between conventionalized and priming-induced repetition. There are good reasons to assume that priming (or, in general, constraints on memory access) is a better model for repetitive language use than mere convention. An analogue to this argument can be found in coordination. Dubey et al. (2005) and Sturt et al. (2010) discuss the conventionalized repetition of syntactic structure in coordinated conjuncts (e.g., phrases coordinated with *and* or *or*), arguing that more general models based on well-known priming effects can account for such constructions as well as more specific models based on a convention to copy syntactic structures. Even if we accepted that repetition is conventionalized, we observe that conventions tend to call for repetition rather than for non-repetition, even in some semantically contrastive use cases. The genesis of such conventions could be explained by the efficiencies of priming-aided language processing.

The empirical difference between the dialogue genres is likely to be functional. It may occur because adaptation *in general* is beneficial to the interlocutor's dialogue goals. It may thus be a *rational* effect (Anderson and Milson, 1989; Chater and Oaksford, 1999). When speakers and listeners performing the Map Task communicated the paths around the landmarks on the map, they were unaware of their present performance, especially in the initial minutes of the interactions, unless the interlocutor pointed out a problem. There was no basis for them on which to actively manipulate their adaptivity. We argue that some Map Task interlocutor pairs were more successful because of their adaptivity, and not vice versa. Based on the IAM's prediction that was confirmed by the empirical, correlational test, we find it is long-term, mechanistic convergence that supports task success.

The IAM is sometimes seen as a theory contrasting the idea that interaction partners establish and track common ground (Clark, 1996). The experiments were designed to examine predictions of the IAM rather than to distinguish the two explanations. However, common ground monitoring cannot explain the effects we observe. If speakers relied solely on feedback monitoring and common ground tracking to optimize their communication, then they would derive no benefit from aligning their syntactic choices. Any correlation between adaptivity and task performance would be unlikely or, at best, be epiphenomenal. Coordination in dialogue (taking place on Clark and Schaefer's (1989) presentational track) may lead to adaptation between speakers based on agreement or grounding of the conveyed message. We argue that such voluntary adaptation affects choices under a speaker's explicit control. Our study examines syntactic adaptation and supports a model where common ground is developed interactively only when needed, but that alignment likely serves human communicators as an inexpensive default strategy.

Conclusion

Our data show that communicative function and linguistic alignment are correlated. Where communication is key, task success is correlated with the syntactic alignment of the two dialogue partners. This finding supports the crucial prediction of the Interactive Alignment Model, which postulates that lower-level alignment such as in speaker's and listener's syntactic choices leads to high-level alignment on a semantic level, improving the dyad's ability to exchange information.

Map Task metric	Correlation with repetition
number of moves	−0.05
(number of yes-no queries) / (number of moves)	−0.12
eye-contact (1/0)	−0.04
familiarity (1/0)	−0.02
experience (num. previous maps)	−0.02

Table 7
Correlation coefficients in Map Task for a count of repeated rules between dialogue halves, normalized by number of rules and residualized for effect of rule frequency.

To our knowledge, this is a first empirical, large-scale test of the model.

As a second theoretical consequence, we need to qualify some of its details of the Interactive Alignment Model. Syntactic priming appears to be moderated by goals or goal-related processes. Priming is more than a mechanistic effect acting on memory retrieval. Implicit access of syntactic rules may be subject to salient cues, e.g., in working memory.

Third, we find no correlation of short-term priming and task success. Conversely, our data show that it is *long-term adaptation* that participates in the alignment cascade at the syntactic level. Interlocutors adapt to or learn from each other. This is a persistent, not transient effect, with lasting consequences for the remainder of the interaction between the two speakers. For a model of communication, we suggest distinguishing between short-term and long-term adaptation effects. Not only may they have different cognitive origins, but they also have different consequences.

The two corpus-based methods we presented let us methodologically quantify structural adaptivity in naturalistic dialogue, at short and long time scales, respectively. The methods apply to general syntactic decisions, connecting to a syntactic theory that can provide a symbolic notion of syntactic choice. There is no need to contrast alternative choices for a given semantics. To provide proof of the generalizability and validity of the priming-task success correlation, we suggested an applied task (estimating task success) and an approach to address it. The task now provides an opportunity to explore and exploit other linguistic and extra-linguistic parameters and connect cognitive psychology to applications.

Acknowledgments

The authors would like to thank Ellen Gurman Bard, Holly Branigan, Alex Fine, Florian Jaeger, Fernanda Ferreira, Dan Jurafsky, Frank Keller, Gregory Mills, Jon Oberlander, Martin Pickering, Frank Ritter and Mark Steedman for discussions throughout the course of this work that helped shape the studies presented.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G.M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R., 1991. The HCRC Map Task corpus. Language and Speech 34, 351–366.

Anderson, A.H., Boyle, E.A., 1993. Forms of introduction in dialogue: their discourse contexts and communicative consequences. Language and Cognitive Processes 9, 101–122.

Anderson, J.R., Bothell, D., Lebiere, C., Matessa, M., 1998. An integrated theory of list memory. Journal of Memory and Language 38, 341–380.

Anderson, J.R., Lebiere, C., 1998. The Atomic Components of Thought. Erlbaum, Mahwah, NJ.

Anderson, J.R., Milson, R., 1989. Human memory: an adaptive perspective. Psychological Review 96, 703–719.

Bard, E.G., Anderson, A.H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., Newlands, A., 2000. Controlling the intelligibility of referring expressions in dialogue. Journal of Memory and Language 42, 1–22.

Bates, D.M., 2014. lme4: Linear mixed-effects models using eigen and s4. URL: `http://cran.r-project.org/web/packages/lme4/`.

Bock, J.K., 1986. Syntactic persistence in language production. Cognitive Psychology 18, 355–387.

Bock, J.K., Dell, G.S., Chang, F., Onishi, K.H., 2007. Persistent structural priming from language comprehension to language production. Cognition 104, 437–458.

Bock, J.K., Griffin, Z.M., 2000. The persistence of structural priming: Transient activation or implicit learning? Journal of Experimental Psychology: General 129, 177.

Bock, J.K., Kroch, A.S., 1989. The isolability of syntactic processing, in: Linguistic structure in language processing. Springer, pp. 157–196.

Bock, J.K., Loebell, H., 1990. Framing sentences. Cognition 35, 1–39.

Boyle, E.A., Anderson, A.H., Newlands, A., 1994. The effects of visibility on dialogue and performance in a cooperative problem solving task. Language and Speech 37, 1–20.

Branigan, H.P., Pickering, M.J., Cleland, A.A., 1999. Syntactic priming in language production: Evidence for rapid decay. Psychonomic Bulletin and Review 6, 635–640.

Branigan, H.P., Pickering, M.J., Cleland, A.A., 2000. Syntactic co-ordination in dialogue. Cognition 75, B13–25.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., 2010. Linguistic alignment between people and computers. Journal of Pragmatics 42, 2355–2368.

Branigan, H.P., Pickering, M.J., Pearson, J., McLean, J.F., Brown, A., 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. Cognition 121, 41 – 57. doi:`10.1016/j.cognition.2011.05.011`.

Brennan, S.E., Hanna, J.E., 2009. Partner-specific adaptation in dialog. Topics in Cognitive Science 1, 274–291.

Carletta, J., Dingare, S., Nissim, M., Nikitina, T., 2004. Using the NITE XML toolkit on the Switchboard corpus to study syntactic choice: A case study, in: Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC, Lisbon, Portugal. pp. 1019–1022.

Chang, F., Dell, G., Bock, J.K., 2006. Becoming syntactic. Psychological Review 113, 234–272.

Chater, N., Oaksford, M., 1999. Ten years of the rational analysis of cognition. Trends in Cognitive Sciences 3, 57–65.

Church, K.W., 2000. Empirial estimates of adaptation: The chance of two Noriegas is closer to $p/2$ than p^2, in: Proceedings of the 18th Conference on Computational Linguistics (COLING), Saarbrücken, Germany. pp. 180–186.

Clark, H.H., 1996. Using Language. Cambridge University Press, Cambridge.

Clark, H.H., Schaefer, E.F., 1989. Contributing to discourse. Cognitive science 13, 259–294.

Dubey, A., Keller, F., Sturt, P., 2005. Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling, in: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing, Vancouver, Canada. pp. 827–834.

Estival, D., 1985. Syntactic priming of the passive in English. Text 5, 7–21.

Ferreira, V., Bock, J.K., 2006. The functions of structural priming. Language and Cognitive Processes 21, 1011–1029.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., Tylén, K., 2012. Coming to terms quantifying the benefits of linguistic coordination. Psychological Science 23, 931–939.

Garrod, S., Anderson, A., 1987. Saying what you mean in dialogue: A study in conceptual and semantic coordination. Cognition 27, 181–218.

Gill, A.J., Harrison, A., Oberlander, J., 2004. Interpersonality: Individual differences and interpersonal priming, in: Proceedings of the 26th Annual Conference of the Cognitive Science Society, Chicago, IL. pp. 464–469.

Gries, S.T., 2005. Syntactic priming: A corpus-based approach. Journal of Psycholinguistic Research 34, 365–399.

Grosz, B.J., Joshi, A.K., Weinstein, S., 1995. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 2, 203–225.

Grosz, B.J., Sidner, C.L., 1986. Attention, intentions, and the structure of discourse. Computational Linguistics 12, 175–204.

Hartsuiker, R.J., Bernolet, S., Schoonbaert, S., Speybroeck, S., Vanderelst, D., 2008. Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. Journal of Memory and Language 58, 214–238.

Healey, P.G.T., 2011. Structural divergence in dialogue, in: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue, Los Angeles, CA. p. 103.

Jaeger, T.F., 2008. Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. Journal of Memory and Language 59, 434–446.

Jonides, J., Lewis, R.L., Nee, D.E., Lustig, C.A., Berman, M.G., Moore, K.S., 2008. The mind and brain of short-term memory. Annual Review of Psychology 59, 193–224.

Kaschak, M.P., Kutta, T.J., Jones, J.L., 2011a. Structural priming as implicit learning: Cumulative priming effects and individual differences. Psychonomic Bulletin & Review 18, 1133–1139.

Kaschak, M.P., Kutta, T.J., Schatschneider, C., 2011b. Long-term cumulative structural priming persists for (at least) one week. Memory & Cognition 39, 381–388.

Kilgarriff, A., 2005. Language is never ever ever random. Corpus Linguistics and Linguistic Theory 1-2, 263–275.

Kucera, H., Francis, W.N., 1967. Computational analysis of present-day American English. Brown University Press, Providence, RI.

Levelt, W.J.M., Kelter, S., 1982. Surface form and memory in question answering. Cognitive Psychology 14, 78–106.

Lewandowsky, S., Duncan, M., Brown, G.D.A., 2004. Time does not cause forgetting in short-term serial recall. Psychonomic Bulletin and Review 11, 771–790.

Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B., 1994. The Penn Treebank: Annotating predicate argument structure, in: Proceedings of the ARPA Human Language Technology Workshop, Plainsboro, NJ. pp. 114–119.

Marr, D., 1982. Vision: A Computational Approach. Freeman & Co, San Francisco, CA.

McKelvie, D., 1998. SDP - Spoken Dialogue Parser. Technical Report HCRC-TR/96. Human Communication Research Centre. Edinburgh, UK.

Pickering, M.J., Branigan, H.P., 1998. The representation of verbs: Evidence from syntactic priming in language production. Journal of Memory and Language 39, 633–651.

Pickering, M.J., Ferreira, V.S., 2008. Structural priming: a critical review. Psychological Bulletin 134, 427.

Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 169–225.

Reitter, D., 2008. Context Effects in Language Production: Models of Syntactic Priming in Dialogue Corpora. Ph.D. thesis. University of Edinburgh.

Reitter, D., Keller, F., 2007. Against sequence priming: Evidence from constituents and distituents in corpus data, in: Proceedings of the 29th Annual Conference of the Cognitive Science Society, Nashville, TN. pp. 1421–1426.

Reitter, D., Keller, F., Moore, J.D., 2011. A computational cognitive model of syntactic priming. Cognitive Science 35, 587–637. doi:http://onlinelibrary.wiley.com/doi/10.1111/j.1551-6709.2010.01165.x/full.

Reitter, D., Moore, J.D., Keller, F., 2006. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation, in: Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci), Vancouver, Canada. pp. 685–690.

Roland, D., Jurafsky, D., 2002. Verb sense and verb subcategorization probabilities, in: Stevenson, S., Merlo, P. (Eds.), The lexical basis of sentence processing: Formal, computational, and experimental issues. John Benjamins, Amsterdam, Netherlands, pp. 325–346.

Scheepers, C., 2003. Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. Cognition 89, 179–205.

Steedman, M., 2000. The Syntactic Process. MIT Press, Cambridge, MA.

Sturt, P., Keller, F., Dubey, A., 2010. Syntactic priming in comprehension: Parallelism effects with and without co-ordination. Journal of Memory and Language 62, 333–351.

Szmrecsanyi, B., 2005. Creatures of habit: A corpus-linguistic analysis of persistence in spoken english. Corpus Linguistics and Linguistic Theory 1, 113–149.

Szmrecsanyi, B., 2006. Morphosyntactic persistence in spoken English. A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis. Mouton de Gruyter, Berlin, Germany.

Appendix

Linear Regression Models

For the short-term priming models, a rule instance *target* is counted as a repetition at distance d iff there is an utterance *prime* which contains the same rule, and *prime* and *target* are exactly d units apart. DIST is the covariate representing the distances d in the data. In most studies presented here, we use *Generalized Linear Mixed Effects Regression Models* (GLMM). GLMMs with a binary response variable can be considered a form of *logistic regression*. The data are assumed to be binomially distributed. We do not generally give classical R^2 figures, as this metric is not appropriate to such GLMMs.

The influence of independent variables, particularly utterance, is expressed as β coefficient in the models: each unit of the independent measure (such as ln(DIST)) adds β to the dependent variable, which is repetition probability expressed in logits. Because effects add up in logit space (i.e., form a *linear* model), they cannot be converted individually to regular probabilities. Interactions such as ln(DIST):PRIMETYPE=CP express the additive contribution to ln(DIST) as a result of another measure. In this case, if PRIMETYPE is "CP" (Comprehension-Production Priming), $\beta_{Dist:PrimeType=CP}$ is added to β_{Dist}. If it is not, nothing is added. Thus, we can contrast the two cases and evaluate its significance by comparing its β value to 0.

Sampling techniques

We draw multiple samples from the same utterance—for several windows at different distances d, but also for each syntactic target rule occurring in the utterance.The resulting dataset has many more non-repetition cases than repetition cases. Balanced sampling addresses the computational problem of fitting the regression models: we include an equal number of data points of repetition and non-repetition cases (PRIME). Conceptually, however, the regression models predict repetition as a function of distance between prime window and target.

Logistic regression model

In short-term priming experiments, we establish priming effects and their interactions with predictors using a logistic regression model (Jaeger, 2008) of the form

$$K(\hat{p}_{Repeated}) = \beta_0 + (\beta_{DIST} + \beta_{DIST:FREQ}\ln(f) + \ldots)\ln(d) + \beta_{FREQ}\ln(f) + \epsilon$$

where K is a logit-link transform, and $\beta_{...}$ are the fitted model parameters. If significantly different from 0, they indicate an effect of the associated main effect (such as prime-target distance d) or interaction (such as between distance and rule frequency). We include a random intercept in our model grouped by target utterance. This declares the several measurements (up to 25 utterances or 15 seconds) as *repeated measurements*, since they depend on the same target rule occurrence and are partially inter-dependent. Interactions with the ln(DIST) main effect estimate the effects of controlled or observed variables with priming.

For all models, we list effects in logits (which can be summed to obtain the final prediction for a given combination of variables). Effects were centered, i.e., categorical variables were coded around a mean of 0. Regression models were estimated using the R package *lme4* (Bates, 2014). In addition to p-values based on standard errors (which we have found to be very similar to p-values obtained via MCMC resampling for our data and models), we report step-wise model ANOVA's obtained with R's *anova* function.

Further possible confounds

Length of the prime window. The response variable used to determine priming encodes whether repetition occurred. Repetition is defined as the occurrence of a given syntactic structure (rule) within a certain time period (*prime window*) as well as at a point later. For short-term priming, this prime window comprises a time period of one second; its distance away from the repetition is varied.

The a-priori probability of repetition occurring anywhere in the prime window also depends on the overall number of rule instances that occur in it. In other words: a fast speaker will show more overall repetition. Our model does not compare repetition rates, but it does depend on decay over time. Whenever speakers slow down their speech production, this may present a possible confound. Then, the a-priori repetition probability is inflated for samples with short prime-target distances, and underestimated for samples with long distances, where the prime window lies outside the peaks. We evaluated the influence of the rule density on priming estimates empirically by fitting additional regression models to both datasets. A correlation of the number of syntactic rule instances and rule repetition was found in Switchboard, but not in Map Task. However, even after controlling for such a correlation, we still found the decay effect that indicates priming.

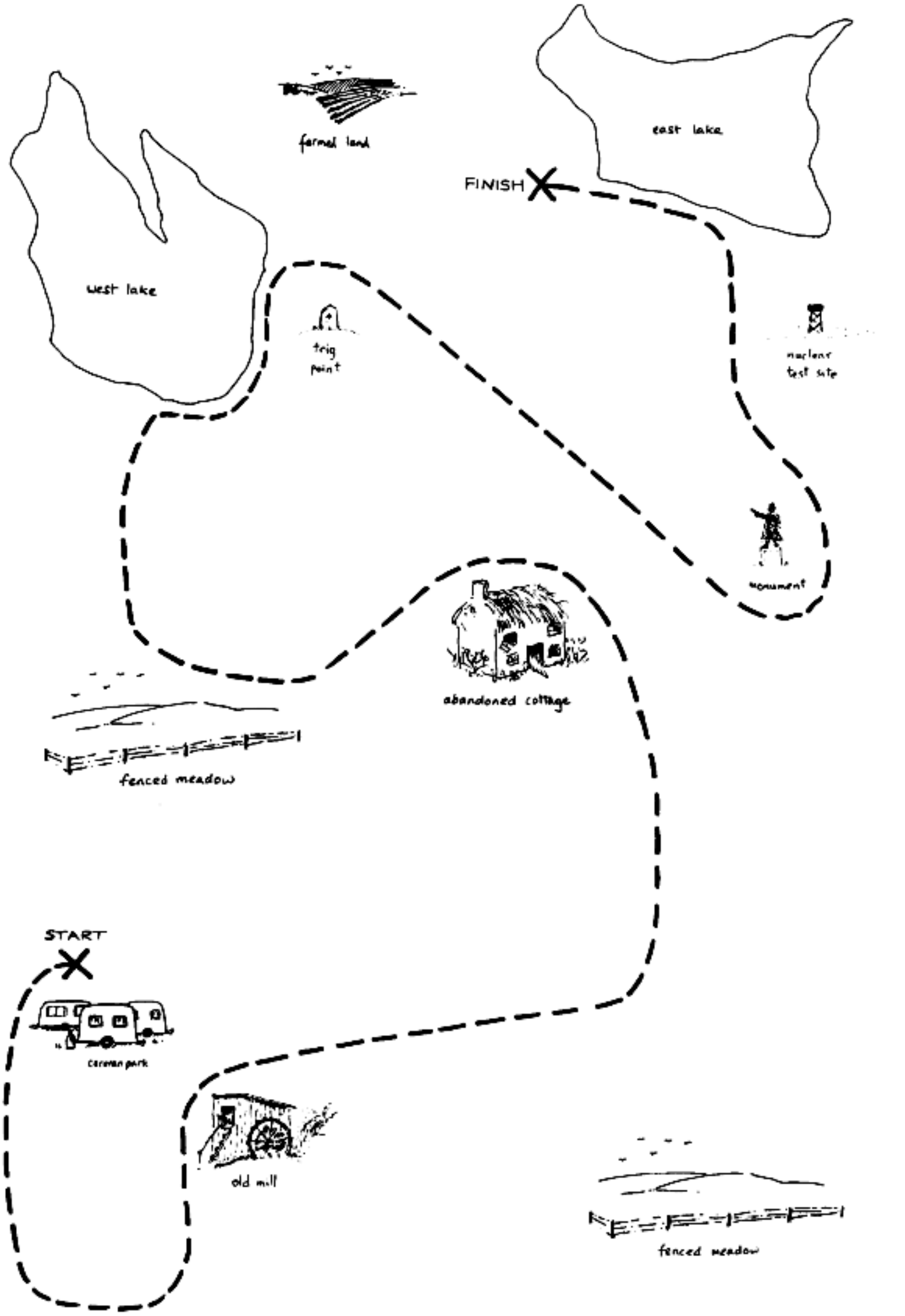


Figure 4
The Map Task corpus: example of a map presented to the Instruction Giver.

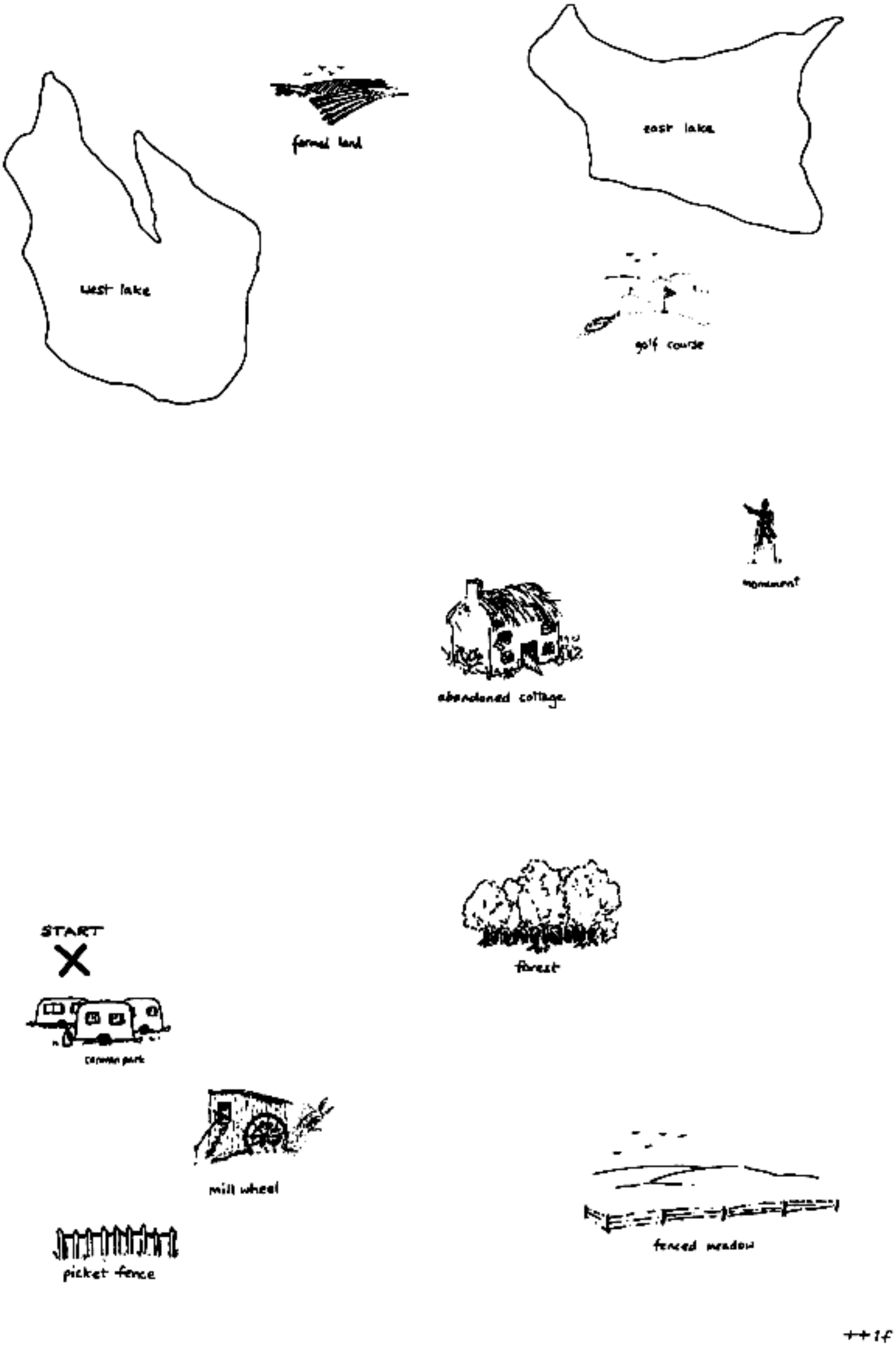


Figure 5
Map from the Map Task corresponding to Fig. 4 used by a subject in the Instruction Follower.