Modeling Semantic Expectation: Using Script Knowledge for Referent Prediction

Ashutosh Modi[1,3] **Ivan Titov**[2,4] **Vera Demberg**[1,3] **Asad Sayeed**[1,3] **Manfred Pinkal**[1,3]
[1] {ashutosh,vera,asayeed,pinkal}@coli.uni-saarland.de
[2] titov@uva.nl
[3] Universität des Saarlandes, Germany
[4] ILLC, University of Amsterdam, the Netherlands

Abstract

Recent research in psycholinguistics has provided increasing evidence that humans predict upcoming content. Prediction also affects perception and might be a key to robustness in human language processing. In this paper, we investigate the factors that affect human prediction by building a computational model that can predict upcoming discourse referents based on linguistic knowledge alone vs. linguistic knowledge jointly with common-sense knowledge in the form of scripts. We find that script knowledge significantly improves model estimates of human predictions. In a second study, we test the highly controversial hypothesis that predictability influences referring expression type but do not find evidence for such an effect.

1 Introduction

Being able to anticipate upcoming content is a core property of human language processing (Kutas et al., 2011; Kuperberg and Jaeger, 2016) that has received a lot of attention in the psycholinguistic literature in recent years. Expectations about upcoming words help humans comprehend language in noisy settings and deal with ungrammatical input. In this paper, we use a computational model to address the question of how different layers of knowledge (linguistic knowledge as well as common-sense knowledge) influence human anticipation.

Here we focus our attention on semantic predictions of *discourse referents* for upcoming noun phrases. This task is particularly interesting because it allows us to separate the semantic task of anticipating an intended referent and the processing of the actual surface form. For example, in the context of *I ordered a medium sirloin steak with fries. Later, the waiter brought* ..., there is a strong expectation of a specific discourse referent, i.e., the referent introduced by the object NP of the preceding sentence, while the possible referring expression could be either *the steak I had ordered*, *the steak*, *our food*, or *it*. Existing models of human prediction are usually formulated using the information-theoretic concept of *surprisal*. In recent work, however, surprisal is usually not computed for DRs, which represent the relevant semantic unit, but for the surface form of the referring expressions, even though there is an increasing amount of literature suggesting that human expectations at different levels of representation have separable effects on prediction and, as a consequence, that the modelling of only one level (the linguistic surface form) is insufficient (Kuperberg and Jaeger, 2016; Kuperberg, 2016; Zarcone et al., 2016). The present model addresses this shortcoming by explicitly modelling and representing common-sense knowledge and conceptually separating the semantic (discourse referent) and the surface level (referring expression) expectations.

Our discourse referent prediction task is related to the NLP task of coreference resolution, but it substantially differs from that task in the following ways: 1) we use only the incrementally available left context, while coreference resolution uses the full text; 2) coreference resolution tries to identify the DR for a given target NP in context, while we look at the expectations of DRs based only on the context

Transactions of the Association for Computational Linguistics, vol. 5, pp. 31–44, 2017. Action Editor: Hwee Tou Ng.
Submission batch: 8/2016 Revision batch: 10/2016; Published 1/2017.

before the target NP is seen.

The distinction between referent prediction and prediction of referring expressions also allows us to study a closely related question in natural language generation: the choice of a type of referring expression based on the predictability of the DR that is intended by the speaker. This part of our work is inspired by a referent guessing experiment by Tily and Piantadosi (2009), who showed that highly predictable referents were more likely to be realized with a pronoun than unpredictable referents, which were more likely to be realized using a full NP. The effect they observe is consistent with a Gricean point of view, or the principle of uniform information density (see Section 5.1). However, Tily and Piantadosi do not provide a computational model for estimating referent predictability. Also, they do not include selectional preference or common-sense knowledge effects in their analysis.

We believe that *script knowledge*, i.e., common-sense knowledge about everyday event sequences, represents a good starting point for modelling conversational anticipation. This type of common-sense knowledge includes temporal structure which is particularly relevant for anticipation in continuous language processing. Furthermore, our approach can build on progress that has been made in recent years in methods for acquiring large-scale script knowledge; see Section 1.1. Our hypothesis is that script knowledge may be a significant factor in human anticipation of discourse referents. Explicitly modelling this knowledge will thus allow us to produce more human-like predictions.

Script knowledge enables our model to generate anticipations about discourse referents that have already been mentioned in the text, as well as anticipations about textually new discourse referents which have been activated due to script knowledge. By modelling event sequences and event participants, our model captures many more long-range dependencies than normal language models are able to. As an example, consider the following two alternative text passages:

We got seated, and had to wait for 20 minutes. Then, the waiter brought the ...

We ordered, and had to wait for 20 minutes. Then, the waiter brought the ...

Preferred candidate referents for the object position of *the waiter brought the ...* are instances of the *food*, *menu*, or *bill* participant types. In the context of the alternative preceding sentences, there is a strong expectation of instances of a *menu* and a *food* participant, respectively.

This paper represents foundational research investigating human language processing. However, it also has the potential for application in assistant technology and embodied agents. The goal is to achieve human-level language comprehension in realistic settings, and in particular to achieve robustness in the face of errors or noise. Explicitly modelling expectations that are driven by common-sense knowledge is an important step in this direction.

In order to be able to investigate the influence of script knowledge on discourse referent expectations, we use a corpus that contains frequent reference to script knowledge, and provides annotations for coreference information, script events and participants (Section 2). In Section 3, we present a large-scale experiment for empirically assessing human expectations on upcoming referents, which allows us to quantify at what points in a text humans have very clear anticipations vs. when they do not. Our goal is to model human expectations, even if they turn out to be incorrect in a specific instance. The experiment was conducted via Mechanical Turk and follows the methodology of Tily and Piantadosi (2009). In section 4, we describe our computational model that represents script knowledge. The model is trained on the gold standard annotations of the corpus, because we assume that human comprehenders usually will have an analysis of the preceding discourse which closely corresponds to the gold standard. We compare the prediction accuracy of this model to human predictions, as well as to two baseline models in Section 4.3. One of them uses only structural linguistic features for predicting referents; the other uses general script-independent selectional preference features. In Section 5, we test whether surprisal (as estimated from human guesses vs. computational models) can predict the type of referring expression used in the original texts in the corpus (pronoun vs. full referring expression). This experiment also has wider implications with respect to the on-going discussion of whether the referring expression choice is dependent on predictability, as predicted by the uniform information density hy-

(I)$^{(1)}_{\text{P\_bather}}$ [**decided**]$_{\text{E\_wash}}$ to take a (bath)$^{(2)}_{\text{P\_bath}}$ yesterday afternoon after working out . Once (I)$^{(1)}_{\text{P\_bather}}$ got back home , (I)$^{(1)}_{\text{P\_bather}}$ [**walked**]$_{\text{E\_enter\_bathroom}}$ to (my)$^{(1)}_{\text{P\_bather}}$ (bathroom)$^{(3)}_{\text{P\_bathroom}}$ and first quickly scrubbed the (bathroom tub)$^{(4)}_{\text{P\_bathtub}}$ by [**turning on**]$_{\text{E\_turn\_water\_on}}$ the (water)$^{(5)}_{\text{P\_water}}$ and rinsing (it)$^{(4)}_{\text{P\_bathtub}}$ clean with a rag . After (I)$^{(1)}_{\text{P\_bather}}$ finished , (I)$^{(1)}_{\text{P\_bather}}$ [**plugged**]$_{\text{E\_close\_drain}}$ the (tub)$^{(4)}_{\text{P\_bathtub}}$ and began [**filling**]$_{\text{E\_fill\_water}}$ (it)$^{(4)}_{\text{P\_bathtub}}$ with warm (water)$^{(5)}_{\text{P\_water}}$ set at about 98 (degrees)$^{(6)}_{\text{P\_temperature}}$.

Figure 1: An excerpt from a story in the InScript corpus. The referring expressions are in parentheses, and the corresponding discourse referent label is given by the superscript. Referring expressions of the same discourse referent have the same color and superscript number. Script-relevant events are in square brackets and colored in orange. Event type is indicated by the corresponding subscript.

pothesis.

The contributions of this paper consist of:

- a large dataset of human expectations, in a variety of texts related to every-day activities.
- an implementation of the conceptual distinction between the semantic level of referent prediction and the type of a referring expression.
- a computational model which significantly improves modelling of human anticipations.
- showing that script knowledge is a significant factor in human expectations.
- testing the hypothesis of Tily and Piantadosi that the choice of the type of referring expression (pronoun or full NP) depends on the predictability of the referent.

1.1 Scripts

Scripts represent knowledge about typical event sequences (Schank and Abelson, 1977), for example the sequence of events happening when eating at a restaurant. Script knowledge thereby includes events like *order*, *bring* and *eat* as well as participants of those events, e.g., *menu*, *waiter*, *food*, *guest*. Existing methods for acquiring script knowledge are based on extracting narrative chains from text (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Jans et al., 2012; Pichotta and Mooney, 2014; Rudinger et al., 2015; Modi, 2016; Ahrendt and Demberg, 2016) or by eliciting script knowledge via Crowdsourcing on Mechanical Turk (Regneri et al., 2010; Frermann et al., 2014; Modi and Titov, 2014).

Modelling anticipated events and participants is motivated by evidence showing that event representations in humans contain information not only about the current event, but also about previous and future states, that is, humans generate anticipations about event sequences during normal language comprehension (Schütz-Bosbach and Prinz, 2007). Script knowledge representations have been shown to be useful in NLP applications for ambiguity resolution during reference resolution (Rahman and Ng, 2012).

2 Data: The InScript Corpus

Ordinary texts, including narratives, encode script structure in a way that is too complex and too implicit at the same time to enable a systematic study of script-based expectation. They contain interleaved references to many different scripts, and they usually refer to single scripts in a point-wise fashion only, relying on the ability of the reader to infer the full event chain using their background knowledge.

We use the InScript corpus (Modi et al., 2016) to study the predictive effect of script knowledge. InScript is a crowdsourced corpus of simple narrative texts. Participants were asked to write about a specific activity (e.g., a restaurant visit, a bus ride, or a grocery shopping event) which they personally experienced, and they were instructed to tell the story as if explaining the activity to a child. This resulted in stories that are centered around a specific scenario and that explicitly mention mundane details. Thus, they generally realize longer event chains associated with a single script, which makes them particularly appropriate to our purpose.

The InScript corpus is labelled with event-type, participant-type, and coreference information. Full verbs are labeled with event type information, heads of all noun phrases with participant types, using scenario-specific lists of event types (such as *enter bathroom, close drain* and *fill water* for the "taking a bath" scenario) and participant types (such as *bather, water* and *bathtub*). On average, each template offers a choice of 20 event types and 18 participant

(I)$^{(1)}$ decided to take a (bath)$^{(2)}$ yesterday afternoon after working out . Once (I)$^{(1)}$ got back home , (I)$^{(1)}$ walked to (my)$^{(1)}$ (bathroom)$^{(3)}$ and first quickly scrubbed the (bathroom tub)$^{(4)}$ by turning on the (water)$^{(5)}$ and rinsing (it)$^{(4)}$ clean with a rag . After (I)$^{(1)}$ finished , (I)$^{(1)}$ plugged **XXXXXX**

Figure 2: An illustration of the Mechanical Turk experiment for the referent cloze task. Workers are supposed to guess the upcoming referent (indicated by **XXXXXX** above). They can either choose from the previously activated referents, or they can write something new.

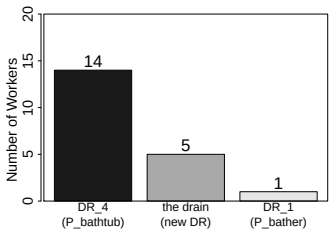

Figure 3: Response of workers corresponding to the story in Fig. 2. Workers guessed two already activated discourse referents (DR) DR_4 and DR_1. Some of the workers also chose the "new" option and wrote different lexical variants of "bathtub drain", a new DR corresponding to the participant type "the drain".

types.

The InScript corpus consists of 910 stories addressing 10 scenarios (about 90 stories per scenario). The corpus has 200,000 words, 12,000 verb instances with event labels, and 44,000 head nouns with participant instances. Modi et al. (2016) report an inter-annotator agreement of 0.64 for event types and 0.77 for participant types (Fleiss' kappa).

We use gold-standard event- and participant-type annotation to study the influence of script knowledge on the expectation of discourse referents. In addition, InScript provides coreference annotation, which makes it possible to keep track of the mentioned discourse referents at each point in the story. We use this information in the computational model of DR prediction and in the DR guessing experiment described in the next section. An example of an annotated InScript story is shown in Figure 1.

3 Referent Cloze Task

We use the InScript corpus to develop computational models for the prediction of discourse referents (DRs) and to evaluate their prediction accuracy. This can be done by testing how often our models manage to reproduce the original discourse referent (cf. also the "narrative cloze" task by (Chambers and Jurafsky, 2008) which tests whether a verb together with a role can be correctly guessed by a model). However, we do not only want to predict the "correct" DRs in a text but also to model human expectation of DRs in context. To empirically assess human expectation, we created an additional database of crowdsourced human predictions of discourse referents in context using Amazon Mechanical Turk. The design of our experiment closely resembles the guessing game of (Tily and Piantadosi, 2009) but extends it in a substantial way.

Workers had to read stories of the InScript corpus [1] and guess upcoming participants: for each target NP, workers were shown the story up to this NP excluding the NP itself, and they were asked to guess the next person or object most likely to be referred to. In case they decided in favour of a discourse referent already mentioned, they had to choose among the available discourse referents by clicking an NP in the preceding text, i.e., some noun with a specific, coreference-indicating color; see Figure 2. Otherwise, they would click the "New" button, and would in turn be asked to give a short description of the new person or object they expected to be mentioned. The percentage of guesses that agree with the actually referred entity was taken as a basis for estimating the surprisal.

The experiment was done for all stories of the test set: 182 stories (20%) of the InScript corpus, evenly taken from all scenarios. Since our focus is on the effect of script knowledge, we only considered those NPs as targets that are direct dependents of script-related events. Guessing started from the third sentence only in order to ensure that a minimum of context information was available. To keep the complexity of the context manageable, we restricted guessing to a maximum of 30 targets and skipped the rest of the story (this applied to 12% of the stories). We collected 20 guesses per NP for 3346 noun phrase instances, which amounts to a total of around 67K guesses. Workers selected a con-

[1] The corpus is available at : `http://www.sfb1102.uni-saarland.de/?page_id=2582`

text NP in 68% of cases and "New" in 32% of cases.

Our leading hypothesis is that script knowledge substantially influences human expectation of discourse referents. The guessing experiment provides a basis to estimate human expectation of already mentioned DRs (the number of clicks on the respective NPs in text). However, we expect that script knowledge has a particularly strong influence in the case of first mentions. Once a script is evoked in a text, we assume that the full script structure, including all participants, is activated and available to the reader.

Tily and Piantadosi (2009) are interested in second mentions only and therefore do not make use of the worker-generated noun phrases classified as "New". To study the effect of activated but not explicitly mentioned participants, we carried out a subsequent annotation step on the worker-generated noun phrases classified as "New". We presented annotators with these noun phrases in their contexts (with co-referring NPs marked by color, as in the M-Turk experiment) and, in addition, displayed all participant types of the relevant script (i.e., the script associated with the text in the InScript corpus). Annotators did not see the "correct" target NP. We asked annotators to either (1) select the participant type instantiated by the NP (if any), (2) label the NP as unrelated to the script, or (3), link the NP to an overt antecedent in the text, in the case that the NP is actually a second mention that had been erroneously labeled as new by the worker. Option (1) provides a basis for a fine-grained estimation of first-mention DRs. Option (3), which we added when we noticed the considerable number of overlooked antecedents, serves as correction of the results of the M-Turk experiment. Out of the 22K annotated "New" cases, 39% were identified as second mentions, 55% were linked to a participant type, and 6% were classified as really novel.

4 Referent Prediction Model

In this section, we describe the model we use to predict upcoming discourse referents (DRs).

4.1 Model

Our model should not only assign probabilities to DRs already explicitly introduced in the preceding text fragment (e.g., "bath" or "bathroom" for the cloze task in Figure 2) but also reserve some probability mass for 'new' DRs, i.e., DRs activated via the script context or completely novel ones not belonging to the script. In principle, different variants of the activation mechanism must be distinguished. For many participant types, a single participant belonging to a specific semantic class is expected (referred to with *the bathtub* or *the soap*). In contrast, the "towel' participant type may activate a set of objects, elements of which then can be referred to with *a towel* or *another towel*. The "bath means" participant type may even activate a group of DRs belonging to different semantic classes (e.g., *bubble bath* and *salts*). Since it is not feasible to enumerate all potential participants, for 'new' DRs we only predict their participant type ("bath means" in our example). In other words, the number of categories in our model is equal to the number of previously introduced DRs plus the number of participant types of the script plus 1, reserved for a new DR not corresponding to any script participant (e.g., *cellphone*). In what follows, we slightly abuse the terminology and refer to all these categories as discourse referents.

Unlike standard co-reference models, which predict co-reference chains relying on the entire document, our model is incremental, that is, when predicting a discourse referent $d^{(t)}$ at a given position t, it can look only in the history $h^{(t)}$ (i.e., the preceding part of the document), excluding the referring expression (RE) for the predicted DR. We also assume that past REs are correctly resolved and assigned to correct participant types (PTs). Typical NLP applications use automatic coreference resolution systems, but since we want to model human behavior, this might be inappropriate, since an automated system would underestimate human performance. This may be a strong assumption, but for reasons explained above, we use gold standard past REs.

We use the following log-linear model ("softmax regression"):

$$p(d^{(t)} = d|h^{(t)}) = \frac{\exp(\mathbf{w}^T\mathbf{f}(d, h^{(t)}))}{\sum_{d'} \exp(\mathbf{w}^T\mathbf{f}(d', h^{(t)}))},$$

where $\mathbf{f}$ is the feature function we will discuss in the following subsection, $\mathbf{w}$ are model parameters, and the summation in the denominator is over the

Feature	Type
Recency	Shallow Linguistic
Frequency	Shallow Linguistic
Grammatical function	Shallow Linguistic
Previous subject	Shallow Linguistic
Previous object	Shallow Linguistic
Previous RE type	Shallow Linguistic
Selectional preferences	Linguistic
Participant type fit	Script
Predicate schemas	Script

Table 1: Summary of feature types

set of categories described above.

Some of the features included in $\mathbf{f}$ are a function of the predicate syntactically governing the unobservable target RE (corresponding to the DR being predicted). However, in our incremental setting, the predicate is not available in the history $h^{(t)}$ for subject NPs. In this case, we use an additional probabilistic model, which estimates the probability of the predicate v given the context $h^{(t)}$, and marginalize out its predictions:

$$p(d^{(t)}=d|h^{(t)})=\sum_v p(v|h^{(t)})\frac{\exp(\mathbf{w}^T\mathbf{f}(d,h^{(t)},v))}{\sum_{d'}\exp(\mathbf{w}^T\mathbf{f}(d',h^{(t)},v))}$$

The predicate probabilities $p(v|h^{(t)})$ are computed based on the sequence of preceding predicates (i.e., ignoring any other words) using the recurrent neural network language model estimated on our training set.[2] The expression $\mathbf{f}(d,h^{(t)},v)$ denotes the feature function computed for the referent d, given the history composed of $h^{(t)}$ and the predicate v.

4.2 Features

Our features encode properties of a DR as well as characterize its compatibility with the context. We face two challenges when designing our features. First, although the sizes of our datasets are respectable from the script annotation perspective, they are too small to learn a richly parameterized model. For many of our features, we address this challenge by using external word embeddings[3] and associate parameters with some simple similarity measures computed using these embeddings. Consequently, there are only a few dozen parameters which need to be estimated from scenario-specific data. Second, in order to test our hypothesis that script information is beneficial for the DR prediction task, we need to disentangle the influence of script information from general linguistic knowledge. We address this by carefully splitting the features apart, even if it prevents us from modeling some interplay between the sources of information. We will describe both classes of features below; also see a summary in Table 1.

4.2.1 Shallow Linguistic Features

These features are based on Tily and Piantadosi (2009). In addition, we consider a selectional preference feature.

Recency feature. This feature captures the distance $l_t(d)$ between the position t and the last occurrence of the candidate DR d. As a distance measure, we use the number of sentences from the last mention and exponentiate this number to make the dependence more extreme; only very recent DRs will receive a noticeable weight: $\exp(-l_t(d))$. This feature is set to 0 for new DRs.

Frequency. The frequency feature indicates the number of times the candidate discourse referent d has been mentioned so far. We do not perform any bucketing.

Grammatical function. This feature encodes the dependency relation assigned to the head word of the last mention of the DR or a special `none` label if the DR is new.

Previous subject indicator. This binary feature indicates whether the candidate DR d is coreferential with the subject of the previous verbal predicate.

Previous object indicator. The same but for the object position.

Previous RE type. This three-valued feature indicates whether the previous mention of the candidate DR d is a pronoun, a non-pronominal noun phrase, or has never been observed before.

4.2.2 Selectional Preferences Feature

The selectional preference feature captures how well the candidate DR d fits a given syntactic position r of a given verbal predicate v. It is computed as the cosine similarity $\mathrm{sim}_{\cos}(\mathbf{x}_d^T,\mathbf{x}_{v,r})$ of a vector-space representation of the DR $\mathbf{x}_d$ and a structured vector-space representation of the pred-

[2] We used RNNLM toolkit (Mikolov et al., 2011; Mikolov et al., 2010) with default settings.

[3] We use 300-dimensional word embeddings estimated on Wikipedia with the skip-gram model of Mikolov et al. (2013): `https://code.google.com/p/word2vec/`

icate $\mathbf{x}_{v,r}$. The similarities are calculated using a Distributional Memory approach similar to that of Baroni and Lenci (2010). Their structured vector space representation has been shown to work well on tasks that evaluate correlation with human thematic fit estimates (Baroni and Lenci, 2010; Baroni et al., 2014; Sayeed et al., 2016) and is thus suited to our task.

The representation $\mathbf{x}_d$ is computed as an average of head word representations of all the previous mentions of DR d, where the word vectors are obtained from the TypeDM model of Baroni and Lenci (2010). This is a count-based, third-order co-occurrence tensor whose indices are a word w_0, a second word w_1, and a complex syntactic relation r, which is used as a stand-in for a semantic link. The values for each (w_0, r, w_1) cell of the tensor are the local mutual information (LMI) estimates obtained from a dependency-parsed combination of large corpora (ukWaC, BNC, and Wikipedia).

Our procedure has some differences with that of Baroni and Lenci. For example, for estimating the fit of an alternative new DR (in other words, $\mathbf{x}_d$ based on no previous mentions), we use an average over head words of all REs in the training set, a "null referent." $\mathbf{x}_{v,r}$ is calculated as the average of the top 20 (by LMI) r-fillers for v in TypeDM; in other words, the prototypical instrument of *rub* may be represented by summing vectors like *towel, soap, eraser, coin...* If the predicate has not yet been encountered (as for subject positions), scores for all scenario-relevant verbs are emitted for marginalization.

4.2.3 Script Features

In this section, we describe features which rely on script information. Our goal will be to show that such common-sense information is beneficial in performing DR prediction. We consider only two script features.

Participant type fit

This feature characterizes how well the participant type (PT) of the candidate DR d fits a specific syntactic role r of the governing predicate v; it can be regarded as a generalization of the selectional preference feature to participant types and also its specialisation to the considered scenario. Given the candidate DR d, its participant type p, and the syntactic relation r, we collect all the predicates in the training set which have the participant type p in the position r. The embedding of the DR $\mathbf{x}_{p,r}$ is given by the average embedding of these predicates. The feature is computed as the dot product of $\mathbf{x}_{p,r}$ and the word embedding of the predicate v.

(I)$^{(1)}$ decided to take a (bath)$^{(2)}$ yesterday afternoon after working out . (I)$^{(1)}$ was getting ready to go out and needed to get cleaned before (I)$^{(1)}$ went so (I)$^{(1)}$ decided to take a (bath)$^{(2)}$. (I)$^{(1)}$ filled the (bathtub)$^{(3)}$ with warm (water)$^{(4)}$ and added some (bubble bath)$^{(5)}$. (I)$^{(1)}$ got undressed and stepped into the (water)$^{(4)}$. (I)$^{(1)}$ grabbed the (soap)$^{(5)}$ and rubbed it on (my)$^{(1)}$ (body)$^{(7)}$ and rinsed **XXXXX**

Figure 4: An example of the referent cloze task. Similar to the Mechanical Turk experiment (Figure 2), our referent prediction model is asked to guess the upcoming DR.

Predicate schemas

The following feature captures a specific aspect of knowledge about prototypical sequences of events. This knowledge is called *predicate schemas* in the recent co-reference modeling work of Peng et al. (2015). In predicate schemas, the goal is to model pairs of events such that if a DR d participated in the first event (in a specific role), it is likely to participate in the second event (again, in a specific role). For example, in the restaurant scenario, if one observes a phrase *John ordered*, one is likely to see *John waited* somewhere later in the document. Specific arguments are not that important (where it is *John* or some other DR), what is important is that the argument is reused across the predicates. This would correspond to the rule *X-subject-of-order* $\rightarrow$ *X-subject-of-eat*.[4] Unlike the previous work, our dataset is small, so we cannot induce these rules directly as there will be very few rules, and the model would not generalize to new data well enough. Instead, we again encode this intuition using similarities in the real-valued embedding space.

Recall that our goal is to compute a feature $\varphi(d, h^{(t)})$ indicating how likely a potential DR d is to follow, given the history $h^{(t)}$. For example, imag-

[4] In this work, we limit ourselves to rules where the syntactic function is the same on both sides of the rule. In other words, we can, in principle, encode the pattern *X pushed Y* $\rightarrow$ *X apologized* but not the pattern *X pushed Y* $\rightarrow$ *Y cried*.

Model Name	Feature Types	Features
Base	Shallow Linguistic Features	Recency, Frequency, Grammatical function, Previous subject, Previous object
Linguistic	Shallow Linguistic Features + Linguistic Feature	Recency, Frequency, Grammatical function, Previous subject, Previous object + Selectional Preferences
Script	Shallow Linguistic Features + Linguistic Feature + Script Features	Recency, Frequency, Grammatical function, Previous subject, Previous object + Selectional Preferences + Participant type fit, Predicate schemas

Table 2: Summary of model features

ine that the model is asked to predict the DR marked by XXXXXX in Figure 4. Predicate-schema rules can only yield previously introduced DRs, so the score $\varphi(d, h^{(t)}) = 0$ for any new DR d. Let us use "soap" as an example of a previously introduced DR and see how the feature is computed. In order to choose which inference rules can be applied to yield "soap", we can inspect Figure 4. There are only two preceding predicates which have DR "soap" as their object (*rubbed* and *grabbed*), resulting in two potential rules *X-object-of-grabbed* $\rightarrow$ *X-object-of-rinsed* and *X-object-of-rubbed* $\rightarrow$ *X-object-of-rinsed*. We define the score $\varphi(d, h^{(t)})$ as the average of the rule scores. More formally, we can write

$$\varphi(d, h^{(t)}) = \frac{1}{|N(d, h^{(t)})|} \sum_{(u,v,r) \in N(d,h^{(t)})} \psi(u, v, r), \quad (1)$$

where $\psi(u, v, r)$ is the score for a rule *X-r-of-u* $\rightarrow$ *X-r-of-v*, $N(d, h^{(t)})$ is the set of applicable rules, and $|N(d, h^{(t)})|$ denotes its cardinality.[5] We define $\varphi(d, h^{(t)})$ as 0, when the set of applicable rules is empty (i.e. $|N(d, h^{(t)})| = 0$).

The scoring function $\psi(u, v, r)$ as a linear function of a joint embedding $\mathbf{x}_{u,v}$ of verbs u and v:

$$\psi(u, v, r) = \alpha_r^T \mathbf{x}_{u,v}.$$

The two remaining questions are (1) how to define the joint embeddings $\mathbf{x}_{u,v}$, and (2) how to estimate the parameter vector α_r. The joint embedding of two predicates, $\mathbf{x}_{u,v}$, can, in principle, be any composition function of embeddings of u and v, for example their sum or component-wise product. Inspired by Bordes et al. (2013), we use the difference between the word embeddings:

$$\psi(u, v, r) = \alpha_r^T (\mathbf{x}_u - \mathbf{x}_v),$$

where $\mathbf{x}_u$ and $\mathbf{x}_v$ are external embeddings of the corresponding verbs. Encoding the succession relation as translation in the embedding space has one desirable property: the scoring function will be largely agnostic to the morphological form of the predicates. For example, the difference between the embeddings of $rinsed$ and $rubbed$ is very similar to that of $rinse$ and rub (Botha and Blunsom, 2014), so the corresponding rules will receive similar scores. Now, we can rewrite the equation (1) as

$$\varphi(d, h^{(t)}) = \alpha_{r(h^{(t)})}^T \frac{\sum_{(u,v,r) \in N(d,h^{(t)})} (\mathbf{x}_u - \mathbf{x}_v)}{|N(d, h^{(t)})|} \quad (2)$$

where $r(h^{(t)})$ denotes the syntactic function corresponding to the DR being predicted (object in our example).

As for the parameter vector α_r, there are again a number of potential ways how it can be estimated. For example, one can train a discriminative classifier to estimate the parameters. However, we opted for a simpler approach—we set it equal to the empirical estimate of the expected feature vector $x_{u,v}$ on the training set:[6]

$$\alpha_r = \frac{1}{D_r} \sum_{l,t} \delta_r(r(h^{(l,t)})) \sum_{(u,v,r') \in N(d^{(l,t)}, h^{(l,t)})} (\mathbf{x}_u - \mathbf{x}_v), \quad (3)$$

where l refers to a document in the training set, t is (as before) a position in the document, $h^{(l,t)}$ and

[5] In all our experiments, rather than considering all potential predicates in the history to instantiate rules, we take into account only 2 preceding verbs. In other words, u and v can be interleaved by at most one verb and $|N(d, h^{(t)})|$ is in $\{0, 1, 2\}$.

[6] This essentially corresponds to using the Naive Bayes model with the simplistic assumption that the score differences are normally distributed with spherical covariance matrices.

Scenario	Human Model		Script Model		Linguistic Model		Tily Model	
	Accuracy	Perplexity	Accuracy	Perplexity	Accuracy	Perplexity	Accuracy	Perplexity
Grocery Shopping	74.80	2.13	68.17	3.16	53.85	6.54	32.89	24.48
Repairing a flat bicycle tyre	78.34	2.72	62.09	3.89	51.26	6.38	29.24	19.08
Riding a public bus	72.19	2.28	64.57	3.67	52.65	6.34	32.78	23.39
Getting a haircut	71.06	2.45	58.82	3.79	42.82	7.11	28.70	15.40
Planting a tree	71.86	2.46	59.32	4.25	47.80	7.31	28.14	24.28
Borrowing book from library	77.49	1.93	64.07	3.55	43.29	8.40	33.33	20.26
Taking Bath	81.29	1.84	67.42	3.14	61.29	4.33	43.23	16.33
Going on a train	70.79	2.39	58.73	4.20	47.62	7.68	30.16	35.11
Baking a cake	76.43	2.16	61.79	5.11	46.40	9.16	24.07	23.67
Flying in an airplane	62.04	3.08	61.31	4.01	48.18	7.27	30.90	30.18
Average	73.63	2.34	62.63	3.88	49.52	7.05	31.34	23.22

Table 3: Accuracies (in %) and perplexities for different models and scenarios. The script model substantially outperforms linguistic and base models (with $p < 0.001$, significance tested with McNemar's test (Everitt, 1992)). As expected, the human prediction model outperforms the script model (with $p < 0.001$, significance tested by McNemar's test).

Model	Accuracy	Perplexity
Linguistic Model	49.52	7.05
Linguistic Model + Predicate Schemas	55.44	5.88
Linguistic Model + Participant type fit	58.88	4.29
Full Script Model (both features)	62.63	3.88

Table 4: Accuracies from ablation experiments.

$d^{(l,t)}$ are the history and the correct DR for this position, respectively. The term $\delta_r(r')$ is the Kronecker delta which equals 1 if $r = r'$ and 0, otherwise. D_r is the total number of rules for the syntactic function r in the training set:

$$D_r = \sum_{l,t} \delta_r(r(h^{(l,t)})) \times |N(d^{(l,t)}, h^{(l,t)})|.$$

Let us illustrate the computation with an example. Imagine that our training set consists of the document in Figure 1, and the trained model is used to predict the upcoming DR in our referent cloze example (Figure 4). The training document includes the pair *X-object-of-scrubbed* $\rightarrow$ *X-object-of-rinsing*, so the corresponding term ($\mathbf{x}_{scrubbed}$ - $\mathbf{x}_{rinsing}$) participates in the summation (3) for α_{obj}. As we rely on external embeddings, which encode semantic similarities between lexical items, the dot product of this term and ($\mathbf{x}_{rubbed}$ - $\mathbf{x}_{rinsed}$) will be high.[7] Consequently, $\varphi(d, h^{(t)})$ is expected to be positive for $d =$ "soap", thus, predicting "soap" as the likely forthcoming DR. Unfortunately, there are other terms $(\mathbf{x_u} - \mathbf{x_v})$ both in expression (3) for α_{obj} and in expression (2) for $\varphi(d, h^{(t)})$. These terms may be irrelevant to the current prediction, as *X-object-of-plugged* $\rightarrow$ *X-object-of-filling* from Figure 1, and may not even encode any valid regularities, as *X-object-of-got* $\rightarrow$ *X-object-of-scrubbed* (again from Figure 1). This may suggest that our feature will be too contaminated with noise to be informative for making predictions. However, recall that independent random vectors in high dimensions are almost orthogonal, and, assuming they are bounded, their dot products are close to zero. Consequently, the products of the relevant ("non-random") terms, in our example ($\mathbf{x}_{scrubbed}$ - $\mathbf{x}_{rinsing}$) and ($\mathbf{x}_{rubbed}$ - $\mathbf{x}_{rinsed}$), are likely to overcome the ("random") noise. As we will see in the ablation studies, the predicate-schema feature is indeed predictive of a DR and contributes to the performance of the full model.

4.3 Experiments

We would like to test whether our model can produce accurate predictions and whether the model's guesses correlate well with human predictions for the referent cloze task.

In order to be able to evaluate the effect of script knowledge on referent predictability, we compare three models: our full *Script model* uses all of the features introduced in section 4.2; the *Linguistic model* relies only on the 'linguistic features' but not the script-specific ones; and the *Base model* includes all the shallow linguistic features. The Base model differs from the linguistic model in that it does not model selectional preferences. Table 2 summarizes features used in different models.

The data set was randomly divided into training (70%), development (10%, 91 stories from 10 sce-

[7] The score would have been even higher, should the predicate be in the morphological form *rinsing* rather than *rinsed*. However, embeddings of *rinsing* and *rinsed* would still be sufficiently close to each other for our argument to hold.

narios), and test (20%, 182 stories from 10 scenarios) sets. The feature weights were learned using L-BFGS (Byrd et al., 1995) to optimize the log-likelihood.

Evaluation against original referents. We calculated the percentage of correct DR predictions. See Table 3 for the averages across 10 scenarios. We can see that the task appears hard for humans: their average performance reaches only 73% accuracy. As expected, the Base model is the weakest system (the accuracy of 31%). Modeling selectional preferences yields an extra 18% in accuracy (Linguistic model). The key finding is that incorporation of script knowledge increases the accuracy by further 13%, although still far behind human performance (62% vs. 73%). Besides accuracy, we use perplexity, which we computed not only for all our models but also for human predictions. This was possible as each task was solved by multiple humans. We used unsmoothed normalized guess frequencies as the probabilities. As we can see from Table 3, the perplexity scores are consistent with the accuracies: the script model again outperforms other methods, and, as expected, all the models are weaker than humans.

As we used two sets of script features, capturing different aspects of script knowledge, we performed extra ablation studies (Table 4). The experiments confirm that both feature sets were beneficial.

Evaluation against human expectations. In the previous subsection, we demonstrated that the incorporation of selectional preferences and, perhaps more interestingly, the integration of automatically acquired script knowledge lead to improved accuracy in predicting discourse referents. Now we turn to another question raised in the introduction: does incorporation of this knowledge make our predictions more human-like? In other words, are we able to accurately estimate human expectations? This includes not only being sufficiently accurate but also making the same kind of incorrect predictions.

In this evaluation, we therefore use human guesses collected during the referent cloze task as our target. We then calculate the relative accuracy of each computational model. As can be seen in Figure 5, the Script model, at approx. 53% accuracy, is a lot more accurate in predicting human guesses than the Linguistic model and the Base model. We can also

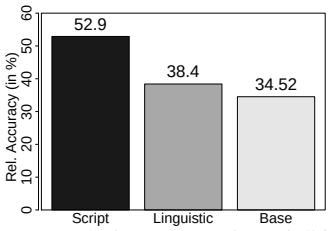

Figure 5: Average relative accuracies of different models w.r.t human predictions.

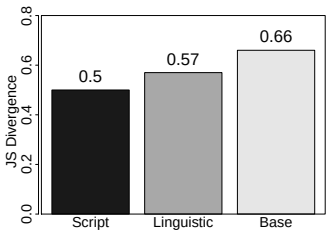

Figure 6: Average Jensen-Shannon divergence between human predictions and models.

observe that the margin between the Script model and the Linguistic model is a lot larger in this evaluation than between the Base model and the Linguistic model. This indicates that the model which has access to script knowledge is much more similar to human prediction behavior in terms of top guesses than the script-agnostic models.

Now we would like to assess if our predictions are similar as distributions rather than only yielding similar top predictions. In order to compare the distributions, we use the Jensen-Shannon divergence (JSD), a symmetrized version of the Kullback-Leibler divergence.

Intuitively, JSD measures the distance between two probability distributions. A smaller JSD value is indicative of more similar distributions. Figure 6 shows that the probability distributions resulting from the Script model are more similar to human predictions than those of the Linguistic and Base models.

In these experiments, we have shown that script knowledge improves predictions of upcoming referents and that the script model is the best among our models in approximating human referent predictions.

5 Referring Expression Type Prediction Model (RE Model)

Using the referent prediction models, we next attempt to replicate Tily and Piantadosi's findings that

the choice of the type of referring expression (pronoun or full NP) depends in part on the predictability of the referent.

5.1 Uniform Information Density hypothesis

The uniform information density (UID) hypothesis suggests that speakers tend to convey information at a uniform rate (Jaeger, 2010). Applied to choice of referring expression type, it would predict that a highly predictable referent should be encoded using a short code (here: a pronoun), while an unpredictable referent should be encoded using a longer form (here: a full NP). Information density is measured using the information-theoretic measure of the *surprisal* S of a message m_i:

$$S(m_i) = -\log P(m_i \mid context)$$

UID has been very successful in explaining a variety of linguistic phenomena; see Jaeger et al. (2016). There is, however, controversy about whether UID affects pronominalization. Tily and Piantadosi (2009) report evidence that writers are more likely to refer using a pronoun or proper name when the referent is easy to guess and use a full NP when readers have less certainty about the upcoming referent; see also Arnold (2001). But other experiments (using highly controlled stimuli) have failed to find an effect of predictability on pronominalization (Stevenson et al., 1994; Fukumura and van Gompel, 2010; Rohde and Kehler, 2014). The present study hence contributes to the debate on whether UID affects referring expression choice.

5.2 A model of Referring Expression Choice

Our goal is to determine whether referent predictability (quantified in terms of surprisal) is correlated with the type of referring expression used in the text. Here we focus on the distinction between pronouns and full noun phrases. Our data also contains a small percentage (ca. 1%) of proper names (like "John"). Due to this small class size and earlier findings that proper nouns behave much like pronouns (Tily and Piantadosi, 2009), we combined pronouns and proper names into a single class of short encodings.

For the referring expression type prediction task, we estimate the surprisal of the referent from each of our computational models from Section 4 as well as the human cloze task. The surprisal of an upcoming discourse referent $d^{(t)}$ based on the previous context $h^{(t)}$ is thereby estimated as:

$$S(d^{(t)}) = -\log\, p(d^{(t)} \mid h^{(t)})$$

In order to determine whether referent predictability has an effect on referring expression type *over and above* other factors that are known to affect the choice of referring expression, we train a logistic regression model with referring expression type as a response variable and discourse referent predictability as well as a large set of other linguistic factors (based on Tily and Piantadosi, 2009) as explanatory variables. The model is defined as follows:

$$p(n^{(t)} = n|d^{(t)}, h^{(t)}) = \frac{\exp(\mathbf{v}^T \mathbf{g}(n, d^t, h^{(t)}))}{\sum_{n'} \exp(\mathbf{v}^T \mathbf{g}(n', d^t, h^{(t)}))},$$

where $d^{(t)}$ and $h^{(t)}$ are defined as before, $\mathbf{g}$ is the feature function, and $\mathbf{v}$ is the vector of model parameters. The summation in the denominator is over NP types (full NP vs. pronoun/proper noun).

5.3 RE Model Experiments

We ran four different logistic regression models. These models all contained exactly the same set of linguistic predictors but differed in the estimates used for referent type surprisal and residual entropy. One logistic regression model used surprisal estimates based on the human referent cloze task, while the three other models used estimates based on the three computational models (Base, Linguistic and Script). For our experiment, we are interested in the choice of referring expression type for those occurrences of references, where a "real choice" is possible. We therefore exclude for our analysis reported below all first mentions as well as all first and second person pronouns (because there is no optionality in how to refer to first or second person). This subset contains 1345 data points.

5.4 Results

The results of all four logistic regression models are shown in Table 5. We first take a look at the results for the linguistic features. While there is a bit of variability in terms of the exact coefficient estimates between the models (this is simply due to small correlations between these predictors and the predictors for surprisal), the effect of all of these features is largely consistent across models. For instance, the positive coefficients for the recency feature means that when a previous mention happened

	Estimate				Std. Error				Pr(>| z |)			
	Human	Script	Linguistic	Base	Human	Script	Linguistic	Base	Human	Script	Linguistic	Base
(Intercept)	-3.4	-3.418	-3.245	-3.061	0.244	0.279	0.321	0.791	<2e-16 ***	<2e-16 ***	<2e-16 ***	0.00011 ***
recency	1.322	1.322	1.324	1.322	0.095	0.095	0.096	0.097	<2e-16 ***	<2e-16 ***	<2e-16 ***	<2e-16 ***
frequency	0.097	0.103	0.112	0.114	0.098	0.097	0.098	0.102	0.317	0.289	0.251	0.262
pastObj	0.407	0.396	0.423	0.395	0.293	0.294	0.295	0.3	0.165	0.178	0.151	0.189
pastSubj	-0.967	-0.973	-0.909	-0.926	0.559	0.564	0.562	0.565	0.0838 .	0.0846 .	0.106	0.101
pastExpPronoun	1.603	1.619	1.616	1.602	0.21	0.207	0.208	0.245	2.19e-14 ***	5.48e-15 ***	7.59e-15 ***	6.11e-11 ***
depTypeSubj	2.939	2.942	2.656	2.417	0.299	0.347	0.429	1.113	<2e-16 ***	<2e-16 ***	5.68e-10 ***	0.02994 *
depTypeObj	1.199	1.227	0.977	0.705	0.248	0.306	0.389	1.109	1.35e-06 ***	6.05e-05 ***	0.0119 *	0.525
surprisal	-0.04	-0.006	0.002	-0.131	0.099	0.097	0.117	0.387	0.684	0.951	0.988	0.735
residualEntropy	-0.009	0.023	-0.141	-0.128	0.088	0.128	0.168	0.258	0.916	0.859	0.401	0.619

Table 5: Coefficients obtained from regression analysis for different models. Two NP types considered: full NP and Pronoun/ProperNoun, with base class full NP. Significance: '***' < 0.001, '**' < 0.01, '*' < 0.05, and '.' < 0.1.

very recently, the referring expression is more likely to be a pronoun (and not a full NP).

The coefficients for the surprisal estimates of the different models are, however, not significantly different from zero. Model comparison shows that they do not improve model fit. We also used the estimated models to predict referring expression type on new data and again found that surprisal estimates from the models did not improve prediction accuracy. This effect even holds for our human cloze data. Hence, it cannot be interpreted as a problem with the models—even human predictability estimates are, for this dataset, not predictive of referring expression type.

We also calculated regression models for the full dataset including first and second person pronouns as well as first mentions (3346 data points). The results for the full dataset are fully consistent with the findings shown in Table 5: there was no significant effect of surprisal on referring expression type.

This result contrasts with the findings by Tily and Piantadosi (2009), who reported a significant effect of surprisal on RE type for their data. In order to replicate their settings as closely as possible, we also included residualEntropy as a predictor in our model (see last predictor in Table 5); however, this did not change the results.

6 Discussion and Future Work

Our study on incrementally predicting discourse referents showed that script knowledge is a highly important factor in determining human discourse expectations. Crucially, the computational modelling approach allowed us to tease apart the different factors that affect human prediction as we cannot manipulate this in humans directly (by asking them to "switch off" their common-sense knowledge).

By modelling common-sense knowledge in terms of event sequences and event participants, our model captures many more long-range dependencies than normal language models. The script knowledge is automatically induced by our model from crowdsourced scenario-specific text collections.

In a second study, we set out to test the hypothesis that uniform information density affects referring expression type. This question is highly controversial in the literature: while Tily and Piantadosi (2009) find a significant effect of surprisal on referring expression type in a corpus study very similar to ours, other studies that use a more tightly controlled experimental approach have not found an effect of predictability on RE type (Stevenson et al., 1994; Fukumura and van Gompel, 2010; Rohde and Kehler, 2014). The present study, while replicating exactly the setting of T&P in terms of features and analysis, did not find support for a UID effect on RE type. The difference in results between T&P 2009 and our results could be due to the different corpora and text sorts that were used; specifically, we would expect that larger predictability effects might be observable at script boundaries, rather than within a script, as is the case in our stories.

A next step in moving our participant prediction model towards NLP applications would be to replicate our modelling results on automatic text-to-script mapping instead of gold-standard data as done here (in order to approximate human level of processing). Furthermore, we aim to move to more complex text types that include reference to several scripts. We plan to consider the recently published ROC Stories corpus (Mostafazadeh et al., 2016), a large crowdsourced collection of topically unrestricted short and simple narratives, as a basis for these next steps in our research.

Acknowledgments

We thank the editors and the anonymous reviewers for their insightful suggestions. We would like to thank Florian Pusse for helping with the Amazon Mechanical Turk experiment. We would also like to thank Simon Ostermann and Tatjana Anikina for helping with the InScript corpus. This research was partially supported by the German Research Foundation (DFG) as part of SFB 1102 'Information Density and Linguistic Encoding', European Research Council (ERC) as part of ERC Starting Grant BroadSem (#678254), the Dutch National Science Foundation as part of NWO VIDI 639.022.518, and the DFG once again as part of the MMCI Cluster of Excellence (EXC 284).

References

Simon Ahrendt and Vera Demberg. 2016. Improving event prediction by representing script participants. In *Proceedings of NAACL-HLT*.

Jennifer E. Arnold. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes*, 31(2):137–162.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of NIPS*.

Jan A. Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *Proceedings of ICML*.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL*.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL*.

Brian S. Everitt. 1992. *The analysis of contingency tables*. CRC Press.

Lea Frermann, Ivan Titov, and Manfred Pinkal. 2014. A hierarchical Bayesian model for unsupervised induction of script knowledge. In *Proceedings of EACL*.

Kumiko Fukumura and Roger P. G. van Gompel. 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language*, 62(1):52–66.

T. Florian Jaeger, Esteban Buz, Eva M. Fernandez, and Helen S. Cairns. 2016. Signal reduction and linguistic encoding. *Handbook of psycholinguistics. Wiley-Blackwell*.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62.

Bram Jans, Steven Bethard, Ivan Vulić, and Marie Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of EACL*.

Gina R. Kuperberg and T. Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59.

Gina R. Kuperberg. 2016. Separate streams or probabilistic inference? What the N400 can tell us about the comprehension of events. *Language, Cognition and Neuroscience*, 31(5):602–616.

Marta Kutas, Katherine A. DeLong, and Nathaniel J. Smith. 2011. A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*.

Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and Jan Cernocky. 2011. RNNLM-recurrent neural network language modeling toolkit. In *Proceedings of the 2011 ASRU Workshop*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Ashutosh Modi and Ivan Titov. 2014. Inducing neural models of script knowledge. *Proceedings of CoNLL*.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. Inscript: Narrative texts annotated with script information. *Proceedings of LREC*.

Ashutosh Modi. 2016. Event embeddings for semantic script modeling. *Proceedings of CoNLL*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. *Proceedings of NAACL*.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of NAACL*.

Karl Pichotta and Raymond J Mooney. 2014. Statistical script learning with multi-argument events. *Proceedings of EACL*.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the Winograd schema challenge. In *Proceedings of EMNLP*.

Michaela Regneri, Alexander Koller, and Manfred Pinkal. 2010. Learning script knowledge with web experiments. In *Proceedings of ACL*.

Hannah Rohde and Andrew Kehler. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927.

Rachel Rudinger, Vera Demberg, Ashutosh Modi, Benjamin Van Durme, and Manfred Pinkal. 2015. Learning to predict script events from domain-specific text. *Proceedings of the International Conference on Lexical and Computational Semantics (*SEM 2015)*.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic fit evaluation: an aspect of selectional preferences. In *Proceedings of the Workshop on Evaluating Vector Space Representations for NLP (RepEval2016)*.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum Associates, Potomac, Maryland.

Simone Schütz-Bosbach and Wolfgang Prinz. 2007. Prospective coding in event representation. *Cognitive processing*, 8(2):93–102.

Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548.

Harry Tily and Steven Piantadosi. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. In *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference*.

Alessandra Zarcone, Marten van Schijndel, Jorrig Vogels, and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing. *Frontiers in Psychology*, 7:844.