

InScript: Narrative texts annotated with script information

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, Manfred Pinkal

Universität des Saarlandes
Saarland, 66123, Germany
{ashutosh, tatianak, simono, pinkal}@coli.uni-saarland.de

Abstract

This paper presents the InScript corpus (Narrative Texts **Instantiating Script** structure). InScript is a corpus of 1,000 stories centered around 10 different scenarios. Verbs and noun phrases are annotated with event and participant types, respectively. Additionally, the text is annotated with coreference information. The corpus shows rich lexical variation and will serve as a unique resource for the study of the role of script knowledge in natural language processing.

Keywords: scripts, narrative texts, script knowledge, common sense knowledge

1. Motivation

A *script* is “a standardized sequence of events that describes some stereotypical human activity such as going to a restaurant or visiting a doctor” (Barr and Feigenbaum, 1981). Script events describe an action/activity along with the involved participants. For example, in the script describing A VISIT TO A RESTAURANT, typical events are ENTERING THE RESTAURANT, ORDERING FOOD or EATING. Participants in this scenario can include animate objects like the WAITER and the CUSTOMER, as well as inanimate objects such as CUTLERY or FOOD.

Script knowledge has been shown to play an important role in text understanding (Cullingford, 1978; Miikkulainen, 1995; Mueller, 2004; Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Modi and Titov, 2014; Rudinger et al., 2015). It guides the expectation of the reader, supports coreference resolution as well as common-sense knowledge inference and enables the appropriate embedding of the current sentence into the larger context. Figure 1 shows the first few sentences of a story describing the scenario TAKING A BATH. Once the TAKING A BATH scenario is evoked by the noun phrase (NP) “a bath”, the reader can effortlessly interpret the definite NP “the faucet” as an implicitly present standard participant of the TAKING A BATH script. Although in this story, “entering the bath room”, “turning on the water” and “filling the tub” are explicitly mentioned, a reader could nevertheless have inferred the “turning on the water” event, even if it was not explicitly mentioned in the text. Table 1 gives an example of typical events and participants for the script describing the scenario TAKING A BATH.

A systematic study of the influence of script knowledge in texts is far from trivial. Typically, text documents (e.g. narrative texts) describing various scenarios evoke many different scripts, making it difficult to study the effect of a single script. Efforts have been made to collect scenario-specific script knowledge via crowdsourcing, for example the OMICS and SMILE corpora (Singh et al., 2002; Regneri et al., 2010; Regneri, 2013), but these corpora describe script events in a pointwise telegram style rather than in full texts.

I was sitting on my couch when I decided that I hadn't taken a bath in a while so I stood up and walked to the bathroom where I turned on the faucet in the sink and began filling the bath with hot water.

While the tub was filling with hot water I put some bubble bath into the stream of hot water coming out of the faucet so that the tubbed filled with not only hot water[...]

Figure 1: An excerpt from a story on the TAKING A BATH script.

This paper presents the InScript¹ corpus (Narrative Texts **Instantiating Script** structure). It is a corpus of simple narrative texts in the form of stories, wherein each story is centered around a specific scenario. The stories have been collected via Amazon Mechanical Turk (M-Turk)². In this experiment, participants (a.k.a. *turkers*) were asked to write down a concrete experience about a bus ride, a grocery shopping event etc. We concentrated on 10 scenarios and collected 100 stories per scenario, giving a total of 1,000 stories with about 200,000 words. Relevant verbs and noun phrases in all stories are annotated with *event types* and *participant types* respectively. Additionally, the texts have been annotated with coreference information in order to facilitate the study of the interdependence between script structure and coreference.

The InScript corpus is a unique resource that provides a basis for studying various aspects of the role of script knowledge in language processing by humans. The acquisition of this corpus is part of a larger research effort that aims at using script knowledge to model the surprisal and information density in written text. Besides InScript, this project also released a corpus of generic descriptions of script activities called DeScript (for **Describing Script** Structure, Wanzare et al. (2016)). DeScript contains a range of short and textually simple phrases that describe script events in the style of OMICS or SMILE (Singh et al., 2002; Regneri et al., 2010). These generic telegram-style descriptions are

¹The corpus can be downloaded at: http://www.sfb1102.uni-saarland.de/?page_id=2582

²<https://www.mturk.com>

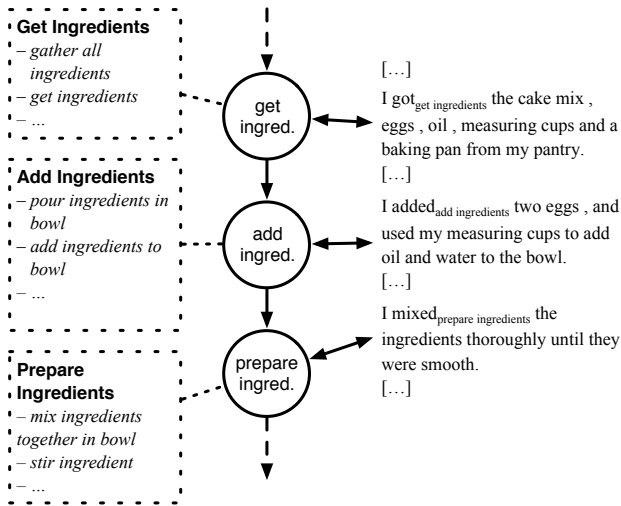


Figure 2: Connecting DeScript and InScript: an example from the BAKING A CAKE scenario (InScript participant annotation is omitted for better readability).

called *Event Descriptions* (EDs); a sequence of such descriptions that cover a complete script is called an *Event Sequence Description* (ESD). Figure 2 shows an excerpt of a script in the BAKING A CAKE scenario. The figure shows event descriptions for 3 different events in the DeScript corpus (left) and fragments of a story in the InScript corpus (right) that instantiate the same event type.

Event types	Participant types
SCREv_TAKE_CLEAN	SCRPART_BATH
_CLOTHES	
SCREv_PREPARE_BATH	SCRPART_BATH_MEANS
SCREv_ENTER_BATHROOM	SCRPART_BATHER
SCREv_TURN_WATER_ON	SCRPART_BATHROOM
SCREv_CHECK_TEMP	SCRPART_BATHTUB
(temperature)	
SCREv_CLOSE_DRAIN	SCRPART_BODY_PART
SCREv_WAIT	SCRPART_CLOTHES
SCREv_TURN_WATER_OFF	SCRPART_DRAIN
SCREv_PUT_BUBBLE	SCRPART_HAIR
_BATH_SCENT	
SCREv_UNDRESS	SCRPART_HAMPER
SCREv_SINK_WATER	SCRPART_IN-BATH
	_ENTERTAINMENT
	(candles, music, books)
SCREv_RELAX	SCRPART_PLUG
SCREv_APPLY_SOAP	SCRPART_SHOWER
	(as bath equipment)
SCREv_WASH	SCRPART_TAP (KNOB)
SCREv_OPEN_DRAIN	SCRPART_TEMPERATURE
SCREv_GET_OUT_BATH	SCRPART_TOWEL
SCREv_GET_TOWEL	SCRPART_WASHING_TOOLS
	(washcloth, soap)
SCREv_DRY	SCRPART_WATER
SCREv_PUT_AFTER_SHOWER	
SCREv_GET_DRESSED	
SCREv_LEAVE	
SCREv_AIR_BATHROOM	

Table 1: Bath scenario template (labels added in the second phase of annotation are marked in bold).

2. Data Collection

2.1. Collection via Amazon M-Turk

We selected 10 scenarios from different available scenario lists (e.g. Regneri et al. (2010), Raisig et al. (2009), and the OMICS corpus (Singh et al., 2002)), including scripts of different complexity (TAKING A BATH vs. FLYING IN AN AIRPLANE) and specificity (RIDING A PUBLIC BUS vs. REPAIRING A FLAT BICYCLE TIRE). For the full scenario list see Table 2.

Texts were collected via the Amazon Mechanical Turk platform, which provides an opportunity to present an online task to humans. In order to gauge the effect of different M-Turk instructions on our task, we first conducted pilot experiments with different variants of instructions explaining the task. We finalized the instructions for the full data collection, asking the turkers to describe a scenario in form of a story *as if explaining it to a child* and to use a minimum of 150 words. The selected instruction variant resulted in comparably simple and explicit scenario-related stories. In the future we plan to collect more complex stories using different instructions. In total 190 turkers participated. All turkers were living in the USA and native speakers of English. We paid USD \$0.50 per story to each turker. On average, the turkers took 9.37 minutes per story with a maximum duration of 17.38 minutes.

2.2. Data Statistics

Statistics for the corpus are given in Table 2. On average, each story has a length of 12 sentences and 217 words with 98 word types on average. Stories are coherent and concentrate mainly on the corresponding scenario. Neglecting auxiliaries, modals and copulas, on average each story has 32 verbs, out of which 58% denote events related to the respective scenario. As can be seen in Table 2, there is some variation in stories across scenarios: The FLYING IN AN AIRPLANE scenario, for example, is most complex in terms of the number of sentences, tokens and word types that are used. This is probably due to the inherent complexity of the scenario: Taking a flight, for example, is more complicated and takes more steps than taking a bath. The average count of sentences, tokens and types is also very high for the BAKING A CAKE scenario. Stories from the scenario often resemble cake recipes, which usually contain very detailed steps, so people tend to give more detailed descriptions in the stories.

For both FLYING IN AN AIRPLANE and BAKING A CAKE, the standard deviation is higher in comparison to other scenarios. This indicates that different turkers described the scenario with a varying degree of detail and can also be seen as an indicator for the complexity of both scenarios. In general, different people tend to describe situations subjectively, with a varying degree of detail.

In contrast, texts from the TAKING A BATH and PLANTING A TREE scenarios contain a relatively smaller number of sentences and fewer word types and tokens. Both planting a tree and taking a bath are simpler activities, which results in generally less complex texts.

The average pairwise word type overlap can be seen as a measure of lexical variety among stories: If it is high, the stories resemble each other more. We can see that stories

Scenario Name	#Stories	Avg. Sentences Per Story	Avg. Word Type Per Story	Avg. Word Count Per Story	Avg. Word Type Overlap
RIDING IN A PUBLIC BUS (BUS)	92	12.3 (4.1)	97.4 (23.3)	215.1 (69.7)	35.7 (7.5)
BAKING A CAKE (CAKE)	97	13.6 (4.7)	102.7 (23.7)	235.5 (78.5)	39.5 (8.1)
TAKING A BATH (BATH)	94	11.5 (2.6)	91.9 (13.1)	197.5 (34.5)	37.9 (6.3)
GOING GROCERY SHOPPING (GROCERY)	95	13.1 (3.7)	102.9 (19.9)	228.3 (58.8)	38.6 (7.8)
FLYING IN AN AIRPLANE (FLIGHT)	86	14.1 (5.6)	113.6 (30.9)	251.2 (99.1)	40.9 (10.3)
GETTING A HAIRCUT (HAIRCUT)	88	13.3 (4.0)	100.6 (19.3)	227.2 (63.4)	39.0 (7.9)
BORROWING A BOOK FROM THE LIBRARY (LIBRARY)	93	11.2 (2.5)	88.0 (14.1)	200.7 (43.5)	34.9 (5.5)
GOING ON A TRAIN (TRAIN)	87	12.3 (3.4)	96.3 (19.2)	210.3 (57.0)	35.3 (6.9)
REPAIRING A FLAT BICYCLE TIRE (BICYCLE)	87	11.4 (3.6)	88.9 (15.0)	203.0 (53.3)	33.8 (5.2)
PLANTING A TREE (TREE)	91	11.0 (3.6)	93.3 (19.2)	201.5 (60.3)	34.0 (6.6)
<i>Average</i>	<i>91</i>	<i>12.4</i>	<i>97.6</i>	<i>216.9</i>	<i>37.0</i>

Table 2: Corpus statistics for different scenarios (standard deviation given in parentheses). The maximum per column is highlighted in **boldface**, the minimum in *boldface italics*.

in the FLYING IN AN AIRPLANE and BAKING A CAKE scenarios have the highest values here, indicating that most turkers used a similar vocabulary in their stories.

In general, the response quality was good. We had to discard 9% of the stories as these lacked the quality we were expecting. In total, we selected 910 stories for annotation.

3. Annotation

This section deals with the annotation of the data. We first describe the final annotation schema. Then, we describe the iterative process of corpus annotation and the refinement of the schema. This refinement was necessary due to the complexity of the annotation.

3.1. Annotation Schema

For each of the scenarios, we designed a specific annotation template. A *script template* consists of scenario-specific event and participant labels. An example of a template is shown in Table 1. All NP heads in the corpus were annotated with a participant label; all verbs were annotated with an event label. For both participants and events, we also offered the label UNCLEAR if the annotator could not assign another label. We additionally annotated coreference chains between NPs. Thus, the process resulted in three layers of annotation: event types, participant types and coreference annotation. These are described in detail below.

Event Type

As a first layer, we annotated event types. There are two kinds of event type labels, scenario-specific event type labels and general labels. The general labels are used across every scenario and mark general features, for example whether an event belongs to the scenario at all. For the scenario-specific labels, we designed a unique template for every scenario, with a list of script-relevant event types that were used as labels. Such labels include for example SCRECV_CLOSE_DRAIN in TAKING A BATH as in Example 1 (see Figure 1 for a complete list for the TAKING A BATH scenario)

- (1) I start by *closing*_{SCRECV_CLOSE_DRAIN} the drain at the bottom of the tub.

The general labels that were used in addition to the script-specific labels in every scenario are listed below:

- SCRECV_OTHER. An event that belongs to the scenario, but its event type occurs too infrequently (for details, see below, Section 3.4.). We used the label “other” because event classification would become too fine-grained otherwise.
Example: After I am dried I put my new clothes on and *clean up*_{SCRECV_OTHER} the bathroom.
- RELNSCRECV. Related non-script event. An event that *can* plausibly happen during the execution of the script and is related to it, but that is not part of the script.
Example: After finding on what I wanted to wear, I went into the bathroom and *shut*_{RELNSCRECV} the door.
- UNRELEV. An event that is unrelated to the script.
Example: I sank into the bubbles and *took*_{UNRELEV} a deep breath.

Additionally, the annotators were asked to annotate verbs and phrases that evoke the script without explicitly referring to a script event with the label EVOKING, as shown in Example 2.

- (2) Today I *took a bath*_{EVOKING} in my new apartment.

Participant Type

As in the case of the event type labels, there are two kinds of participant labels: general labels and scenario-specific labels. The latter are part of the scenario-specific templates, e.g. SCRPART_DRAIN in the TAKING A BATH scenario, as can be seen in Example 3.

- (3) I start by closing the *drain*_{SCRPART_DRAIN} at the bottom of the tub.

The general labels that are used across all scenarios mark noun phrases with scenario-independent features. There are the following general labels:

- SCRPART_OTHER. A participant that belongs to the scenario, but its participant type occurs only infrequently.

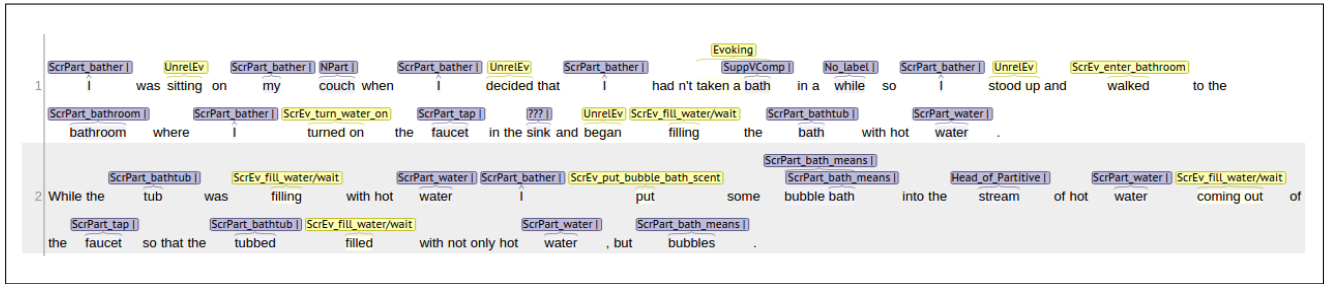


Figure 3: Sample event and participant annotation for the TAKING A BATH script.

Example: I find my *bath mat*_{SCRPART_OTHER} and lay it on the floor to keep the floor dry.

- NPART. Non-participant. A referential NP that does not belong to the scenario.
Example: I washed myself carefully because I did not want to spill water onto the *floor*_{NPART}.
- SUPPVCOMP. A support verb complement. For further discussion of this label, see Section 3.5.
Example: I sank into the bubbles and took a deep *breath*_{SUPPVCOMP}.
- HEAD_OF_PARTITIVE. The head of a partitive or a partitive-like construction. For a further discussion of this label cf. Section 3.5.
Example: I grabbed a *bar*_{HEAD_OF_PARTITIVE} of soap and lathered my body.
- NO_LABEL. A non-referential noun phrase that cannot be labeled with another label.
Example: I sat for a *moment*_{NO_LABEL}, relaxing, allowing the warm water to sooth my skin.

All NPs labeled with one of the labels SUPPVCOMP, HEAD_OF_PARTITIVE or NO_LABEL are considered to be non-referential. NO_LABEL is used mainly in four cases in our data: non-referential time expressions (*in a while*, *a million times better*), idioms (*no matter what*), the non-referential “it” (*it felt amazing*, *it is better*) and other abstracta (*a lot better*, *a little bit*).

In the first annotation phase, annotators were asked to mark verbs and noun phrases that have an event or participant type, that is *not* listed in the template, as MISSSCREVEV/MISSSCRPART (missing script event or participant, resp.). These annotations were used as a basis for extending the templates (see Section 3.4.) and replaced later by newly introduced labels or SCREVEV_OTHER and SCRPART_OTHER respectively.

Coreference Annotations

All noun phrases were annotated with coreference information indicating which entities denote the same discourse referent. The annotation was done by linking heads of NPs (see Example 4, where the links are indicated by coindexing). As a rule, we assume that each element of a coreference chain is marked with the same participant type label.

- (4) I_{COREF1} washed my_{COREF1} entire $body_{COREF2}$, starting with my_{COREF1} $face_{COREF3}$ and ending with the

$toes_{COREF4}$. I_{COREF1} always wash my_{COREF1} $toes_{COREF4}$ very thoroughly ...

The assignment of an entity to a referent is not always trivial, as is shown in Example 5. There are some cases in which two discourse referents are grouped in a plural NP. In the example, *those things* refers to the group made up of *shampoo*, *soap* and *sponge*. In this case, we asked annotators to introduce a new coreference label, the name of which indicates which referents are grouped together (COREF_GROUP_WASHING_TOOLS). All NPs are then connected to the group phrase, resulting in an additional coreference chain.

- (5) I_{COREF1} made sure that I_{COREF1} have my_{COREF1} *shampoo*_{COREF2 + COREF_GROUP_WASHING_TOOLS}, *soap*_{COREF3 + COREF_GROUP_WASHING_TOOLS} and *sponge*_{COREF4 + COREF_GROUP_WASHING_TOOLS} ready to get in. Once I_{COREF1} have those *things*_{COREF_GROUP_WASHING_TOOLS} I_{COREF1} sink into the bath. ... I_{COREF1} applied some *soap*_{COREF3} on my_{COREF1} body and used the *sponge*_{COREF4} to scrub a bit. ... I_{COREF1} rinsed the *shampoo*_{COREF2}.

Example 5 thus contains the following coreference chains:

- (6) COREF1: $I \rightarrow I \rightarrow my \rightarrow I \rightarrow I \rightarrow I \rightarrow my \rightarrow I$
 COREF2: shampoo \rightarrow shampoo
 COREF3: soap \rightarrow soap
 COREF4: sponge \rightarrow sponge
 COREF_GROUP_WASHING_TOOLS: shampoo \rightarrow soap \rightarrow sponge \rightarrow things

3.2. Development of the Schema

The templates were carefully designed in an iterated process. For each scenario, one of the authors of this paper provided a preliminary version of the template based on the inspection of some of the stories; for a subset of the scenarios, preliminary templates developed at our department for a psycholinguistic experiment on script knowledge were used as a starting point. Subsequently, the authors manually annotated 5 randomly selected texts for each of the scenarios based on the preliminary template. Necessary extensions and changes in the templates were discussed and agreed upon. Most of the cases of disagreement were related to the granularity of the event and participant types. We agreed on the *script-specific functional equivalence* as a guiding principle. For example, reading a book, listening to music and having a conversation are subsumed under

the same event label in the FLIGHT scenario, because they have the common function of in-flight entertainment in the scenario. In contrast, we assumed different labels for the cake tin on the one hand and other utensils (bowls etc.) on the other hand, since they have different functions in the BAKING A CAKE scenario and accordingly occur with different script events.

Note that scripts and templates as such are not meant to describe an activity as exhaustively as possible and to mention all steps that are logically necessary. Instead, scripts describe cognitively prominent events in an activity. An example can be found in the FLIGHT scenario. While more than a third of the turkers mentioned the event of fastening the seat belts in the plane (BUCKLE_SEAT_BELT), no person wrote about *undoing* their seat belts again, although in reality both events appear equally often. Consequently, we added an event type label for buckling up, but no label for undoing the seat belts.

3.3. First Annotation Phase

We used the WebAnno annotation tool (Yimam et al., 2013) for our project. The stories from each scenario were distributed among four different annotators. In a calibration phase, annotators were presented with some sample texts for test annotations; the results were discussed with the authors. Throughout the whole annotation phase, annotators could discuss any emerging issues with the authors. All annotations were done by undergraduate students of computational linguistics. The annotation was rather time-consuming due to the complexity of the task, and thus we decided for single annotation mode. To assess annotation quality, a small sample of texts was annotated by all four annotators and their inter-annotator agreement was measured (see Section 4.1.). It was found to be sufficiently high.

Annotation of the corpus together with some pre- and post-processing of the data required about 500 hours of work. All stories were annotated with event and participant types (a total of 12,188 and 43,946 instances, respectively). On average there were 7 coreference chains per story with an average length of 6 tokens.

3.4. Modification of the Schema

After the first annotation round, we extended and changed the templates based on the results. As mentioned before, we used MISSSCREVEV and MISSSCRPART labels to mark verbs and noun phrases instantiating events and participants for which no appropriate labels were available in the templates. Based on the instances with these labels (a total of 941 and 1717 instances, respectively), we extended the guidelines to cover the sufficiently frequent cases.

In order to include new labels for event and participant types, we tried to estimate the number of instances that would fall under a certain label. We added new labels according to the following conditions:

- For the participant annotations, we added new labels for types that we expected to appear at least 10 times in total in at least 5 different stories (i.e. in approximately 5% of the stories).

- For the event annotations, we chose those new labels for event types that would appear in at least 5 different stories.

In order to avoid too fine a granularity of the templates, all other instances of MISSSCREVEV and MISSSCRPART were re-labeled with SCREVEV_OTHER and SCRPART_OTHER. We also relabeled participants and events from the first annotation phase with SCREVEV_OTHER and SCRPART_OTHER, if they did not meet the frequency requirements. The event label AIR_BATHROOM (the event of letting fresh air into the room after the bath), for example, was only used once in the stories, so we relabeled that instance to SCREVEV_OTHER.

Additionally, we looked at the DeScript corpus (Wanzare et al., 2016), which contains manually clustered event phrase sets for the 10 scenarios that are also covered by InScript (see Section 4.3.). Every such set contains event descriptions that describe a certain event type. We extended our templates with additional labels for these events, if they were not yet part of the template.

3.5. Special Cases

Noun-Noun Compounds. Noun-noun compounds were annotated twice with the same label (whole span plus the head noun), as indicated by Example 7. This redundant double annotation is motivated by potential processing requirements.

(7) I get my (*wash* (*cloth* _{SCRPART_WASHING_TOOLS})),
_{SCRPART_WASHING_TOOLS} and put it under the water.

Support Verb Complements. A special treatment was given to support verb constructions such as *take time*, *get home* or *take a seat* in Example 8. The semantics of the verb itself is highly underspecified in such constructions; the event type is largely dependent on the object NP. As shown in Example 8, we annotate the head verb with the event type described by the whole construction and label its object with SUPPVCOMP (support verb complement), indicating that it does not have a proper reference.

(8) I step into the tub and *take*_{SCREVEV_SINK_WATER} a
*seat*_{SUPPVCOMP}.

Head of Partitive. We used the HEAD_OF_PARTITIVE label for the heads in partitive constructions, assuming that the only referential part of the construction is the complement. This is not completely correct, since different partitive heads vary in their degree of concreteness (cf. Examples 9 and 10), but we did not see a way to make the distinction sufficiently transparent to the annotators.

(9) Our seats were at the *back*_{HEAD_OF_PARTITIVE} of the
*train*_{SCRPART_TRAIN}.

(10) In the library you can always find a *cou-
 ple*_{HEAD_OF_PARTITIVE} of interesting books_{SCRPART_BOOK}.

Scenario	Average Fleiss' Kappa			
	All Labels		Script Labels	
	Events	Participants	Events	Participants
BUS	0.68	0.74	0.76	0.74
CAKE	0.61	0.76	0.64	0.75
FLIGHT	0.65	0.70	0.62	0.69
GROCERY	0.64	0.80	0.73	0.80
HAIRCUT	0.64	0.84	0.67	0.86
TREE	0.59	0.76	0.63	0.76
<i>Average</i>	<i>0.64</i>	<i>0.77</i>	<i>0.68</i>	<i>0.77</i>

(a) Average Fleiss' Kappa.

Scenario	%Coreference Agreement
BUS	88.9
CAKE	94.7
FLIGHT	93.6
GROCERY	93.4
HAIRCUT	94.3
TREE	78.3
<i>Average</i>	<i>90.5</i>

(b) Coreference agreement.

Figure 4: Inter-annotator agreement statistics.

Mixed Participant Types. Group denoting NPs sometimes refer to groups whose members are instances of different participant types. In Example 11, the first-person plural pronoun refers to the group consisting of the passenger (*I*) and a non-participant (*my friend*). To avoid a proliferation of event type labels, we labeled these cases with UNCLEAR.

- (11) $I_{\text{SCR_PART_PASSENGER}}$ wanted to visit $my_{\text{SCR_PART_PASSENGER}}$ $friend_{\text{NPART}}$ in New York. ... We_{UNCLEAR} met at the train station.

We made an exception for the GETTING A HAIRCUT scenario, where the mixed participant group consisting of the hairdresser and the customer occurs very often, as in Example 11. Here, we introduced the additional ad-hoc participant label SCR_PART_HAIRDRESSER_CUSTOMER.

- (12) While $Susan_{\text{SCR_PART_HAIRDRESSER}}$ is cutting $my_{\text{SCR_PART_CUSTOMER}}$ hair $we_{\text{SCR_PART_HAIRDRESSER_CUSTOMER}}$ usually talk a bit.

4. Data Analysis

4.1. Inter-Annotator Agreement

In order to calculate inter-annotator agreement, a total of 30 stories from 6 scenarios were randomly chosen for parallel annotation by all 4 annotators after the first annotation phase³. We checked the agreement on these data using Fleiss' Kappa (Fleiss, 1971). The results are shown in Figure 4a and indicate moderate to substantial agreement (Landis and Koch, 1977). Interestingly, if we calculated the Kappa only on the subset of cases that were annotated with script-specific event and participant labels by all annotators, results were better than those of the evaluation on all labeled instances (including also unrelated and related non-script events). This indicates one of the challenges of the annotation task: In many cases it is difficult to decide whether a particular event should be considered a central script event, or an event loosely related or unrelated to the script.

For coreference chain annotation, we calculated the percentage of pairs which were annotated by at least 3 annotators (qualified majority vote) compared to the set of those

³We did not test for inter-annotator agreement after the second phase, since we did not expect the agreement to change drastically due to the only slight changes in the annotation schema.

pairs annotated by at least one person (see Figure 4b). We take the result of 90.5% between annotators to be a good agreement.

4.2. Annotated Corpus Statistics

Scenario	Events	Participants
BATH	20	18
BICYCLE	16	16
BUS	17	17
CAKE	19	17
FLIGHT	29	26
GROCERY	19	18
HAIRCUT	26	24
LIBRARY	17	18
TRAIN	15	20
TREE	14	15
<i>Average</i>	<i>19.2</i>	<i>18.9</i>

Figure 5: The number of participants and events in the templates.

Figure 5 gives an overview of the number of event and participant types provided in the templates. TAKING A FLIGHT and GETTING A HAIRCUT stand out with a large number of both event and participant types, which is due to the inherent complexity of the scenarios. In contrast, PLANTING A TREE and GOING ON A TRAIN contain the fewest labels. There are 19 event and participant types on average.

	avg	min	max
event annotations in a story	15.9	1	52
event types in a story	10.1	1	23
participant annotations in a story	52.3	16	164
participant types in a story	10.9	2	25
coref chains	7.3	0	23
tokens per chain	6	2	52

Figure 6: Annotation statistics over all scenarios.

Figure 6 presents overview statistics about the usage of event labels, participant labels and coreference chain annotations. As can be seen, there are usually many more mentions of participants than events. For coreference chains, there are some chains that are really long (which also results in a large scenario-wise standard deviation). These chains usually describe the protagonist.

We also found again that the FLYING IN AN AIRPLANE

scenario stands out in terms of participant mentions, event mentions and average number of coreference chains.

Figure 7 shows for every participant label in the BAKING A CAKE scenario the number of stories which they occurred in. This indicates how relevant a participant is for the script. As can be seen, a small number of participants are highly prominent: COOK, INGREDIENTS and CAKE are mentioned in every story. The fact that the protagonist appears most often consistently holds for all other scenarios, where the acting person appears in every story, and is mentioned most frequently.

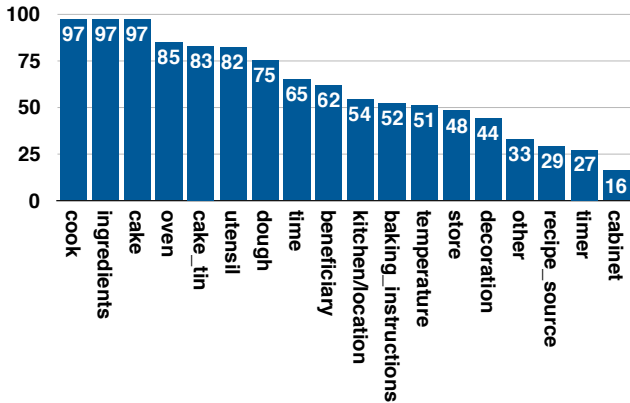


Figure 7: The number of stories in the BAKING A CAKE scenario that contain a certain participant label.

Figure 8 shows the distribution of participant/event type labels over all appearances over all scenarios on average. The groups stand for the most frequently appearing label, the top 2 to 5 labels in terms of frequency and the top 6 to 10. SCREV_OTHER and SCRPART_OTHER are shown separately. As can be seen, the most frequently used participant label (the protagonist) makes up about 40% of overall participant instances. The four labels that follow the protagonist in terms of frequency together appear in 37% of the cases. More than 2 out of 3 participants in total belong to one of only 5 labels.

In contrast, the distribution for events is more balanced. 13% of all event instances have the most prominent event type.

SCREV_OTHER and SCRPART_OTHER both appear as labels in at most 5% of all event and participant instantiations: The specific event and participant type labels in our templates cover by far most of the instances.

In Figure 9, we grouped participants similarly into the first, the top 2-5 and top 6-10 most frequently appearing participant types. The figure shows for each of these groups the average frequency per story, and in the rightmost column the overall average. The results correspond to the findings from the last paragraph.

4.3. Comparison to the DeScript Corpus

As mentioned previously, the InScript corpus is part of a larger research project, in which also a corpus of a different kind, the DeScript corpus, was created. DeScript covers 40 scenarios, and also contains the 10 scenarios from InScript. This corpus contains texts that describe scripts on

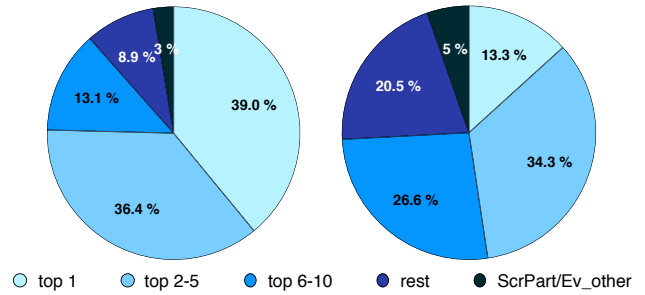


Figure 8: Distribution of participants (left) and events (right) for the 1, the top 2-5, top 6-10 most frequently appearing events/participants, SCREV_OTHER and SCRPART_OTHER and the rest.

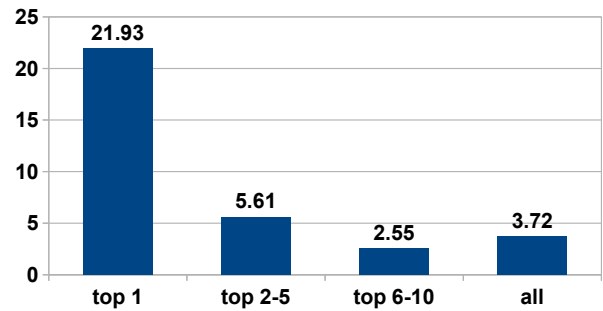


Figure 9: Average number of participant mentions for a story, for the first, the top 2-5, top 6-10 most frequently appearing events/participants, and the overall average.

an abstract, generic level, while InScript contains instantiations of scripts in narrative texts. Script events in DeScript are described in a very simple, telegram-style language (see Figure 2). Since one of the long-term goals of the project is to align the InScript texts with the script structure given from DeScript, it is interesting to compare both resources. The InScript corpus exhibits much more lexical variation than DeScript. Many approaches use the *type-token ratio* to measure this variance. However, this measure is known to be sensitive to text length (see e.g. Tweedie and Baayen (1998)), which would result in very small values for InScript and relatively large ones for DeScript, given the large average difference of text lengths between the corpora. Instead, we decided to use the *Measure of Textual Lexical Diversity (MTLD)* (McCarthy and Jarvis, 2010; McCarthy, 2005), which is familiar in corpus linguistics. This metric measures the average number of tokens in a text that are needed to retain a type-token ratio above a certain threshold. If the *MTLD* for a text is high, many tokens are needed to lower the type-token ratio under the threshold, so the text is lexically diverse. In contrast, a low *MTLD* indicates that only a few words are needed to make the type-token ratio drop, so the lexical diversity is smaller. We use the threshold of 0.71, which is proposed by the authors as a well-proven value.

Figure 10 compares the lexical diversity of both resources. As can be seen, the InScript corpus with its narrative texts is generally much more diverse than the DeScript corpus with its short event descriptions, across all scenarios. For both resources, the FLYING IN AN AIRPLANE scenario is

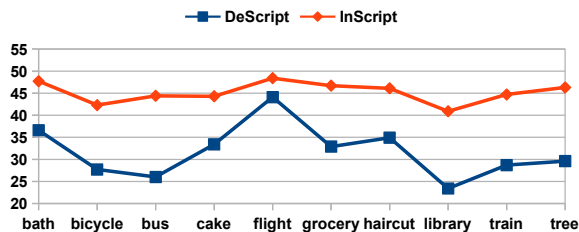


Figure 10: MTLD values for DeScript and InScript, per scenario.

most diverse (as was also indicated above by the mean word type overlap). However, the difference in the variation of lexical variance of scenarios is larger for DeScript than for InScript. Thus, the properties of a scenario apparently influence the lexical variance of the event descriptions more than the variance of the narrative texts.

We used entropy (Shannon, 1948) over lemmas to measure the variance of lexical realizations for events. We excluded events for which there were less than 10 occurrences in DeScript or InScript. Since there is only an event annotation for 50 ESDs per scenario in DeScript, we randomly sampled 50 texts from InScript for computing the entropy to make the numbers more comparable.

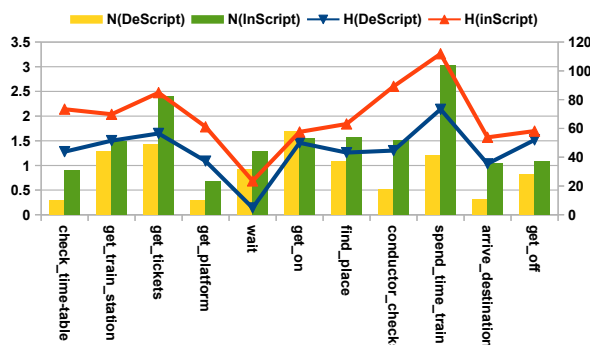


Figure 11: Entropy over verb lemmas for events (left y-axis, $H(x)$) in the GOING ON A TRAIN SCENARIO. Bars in the background indicate the absolute number of occurrence of instances (right y-axis, $N(x)$).

Figure 11 shows as an example the entropy values (abbreviated as H) for the event types in the GOING ON A TRAIN scenario. As can be seen in the graph, the entropy for InScript is in general higher than for DeScript. In the stories, a wider variety of verbs is used to describe events. There are also large differences between events: While WAIT has a really low entropy, SPEND_TIME_TRAIN has an extremely high entropy value. This event type covers many different activities such as reading, sleeping etc.

5. Conclusion

In this paper we described the InScript corpus of 1,000 narrative texts annotated with script structure and coreference information. We described the annotation process, various difficulties encountered during annotation and different remedies that were taken to overcome these. One of the future research goals of our project is also concerned

with finding automatic methods for text-to-script mapping, i.e. for the alignment of text segments with script states. We consider InScript and DeScript together as a resource for studying this alignment. The corpus shows rich lexical variation and will serve as a unique resource for the study of the role of script knowledge in natural language processing.

Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 'Information Density and Linguistic Encoding'.

6. References

- Barr, A. and Feigenbaum, E. A. (1981). *The Handbook of Artificial Intelligence*. Addison-Wesley.
- Chambers, N. and Jurafsky, D. (2008). Unsupervised learning of narrative event chains. *Proceedings of ACL-08*.
- Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*.
- Cullingford, R. E. (1978). Script application: Computer understanding of newspaper stories. Technical report, DTIC Document.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):pp. 159–174.
- McCarthy, P. M. and Jarvis, S. (2010). MtlD, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*.
- McCarthy, P. M. (2005). *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)*. Ph.D. thesis, The University of Memphis.
- Miikkulainen, R. (1995). Script-based inference and memory retrieval in subsymbolic story processing. *Applied Intelligence*, 5(2):137–163.
- Modi, A. and Titov, I. (2014). Inducing neural models of script knowledge. In *CoNLL*, volume 14, pages 49–57.
- Mueller, E. T. (2004). Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research*, 5(4):307–340.
- Raisig, S., Welke, T., Hagendorf, H., and Van Der Meer, E. (2009). Insights into knowledge representation: The influence of amodal and perceptual variables on event knowledge retrieval from memory. *Cognitive Science*, 33(7):1252–1266.
- Regneri, M., Koller, A., and Pinkal, M. (2010). Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 979–988, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Regneri, M. (2013). *Event Structures in Knowledge, Pictures and Text*. Ph.D. thesis, Universität des Saarlandes.

- Rudinger, R., Demberg, V., Modi, A., Van Durme, B., and Pinkal, M. (2015). Learning to predict script events from domain-specific text. *Lexical and Computational Semantics (*SEM 2015)*, page 205.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423.
- Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In *On the move to meaningful internet systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer.
- Tweedie, F. J. and Baayen, R. H. (1998). How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5):323–352.
- Wanzare, L. D. A., Zarccone, A., Thater, S., and Pinkal, M. (2016). A crowdsourced database of event sequence descriptions for the acquisition of high-quality script knowledge. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.
- Yimam, S. M., Gurevych, I., de Castilho, R. E., and Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *ACL (Conference System Demonstrations)*, pages 1–6.