DeScript: A Crowdsourced Corpus for the Acquisition of High-Quality Script Knowledge

Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, Manfred Pinkal
Universität des Saarlandes
Saarland, 66123, Germany
wanzare@coli.uni-saarland.de, zarcone@coli.uni-saarland.de, stth@coli.uni-saarland.de, pinkal@coli.uni-saarland.de

Abstract

Scripts are standardized event sequences describing typical everyday activities, which play an important role in the computational modeling of cognitive abilities (in particular for natural language processing). We present a large-scale crowdsourced collection of explicit linguistic descriptions of script-specific event sequences (40 scenarios with 100 sequences each). The corpus is enriched with crowdsourced alignment annotation on a subset of the event descriptions, to be used in future work as seed data for automatic alignment of event descriptions (for example via clustering). The event descriptions to be aligned were chosen among those expected to have the strongest corrective effect on the clustering algorithm. The alignment annotation was evaluated against a gold standard of expert annotators. The resulting database of partially-aligned script-event descriptions provides a sound empirical basis for inducing high-quality script knowledge, as well as for any task involving alignment and paraphrase detection of events.

Keywords: scripts, events, crowdsourcing, paraphrase

1. Introduction

When carrying out everyday activities, having a conversation, watching movies or reading novels or newspapers, we make abundant use of script knowledge, that is, knowledge of standardized event sequences describing typical everyday activities, such as *baking a cake* or *eating in a fast food restaurant* (Schank and Abelson, 1977; Barr and Feigenbaum, 1981). Script knowledge plays an important role for the computational modeling of cognitive abilities (in particular for natural language processing), but making this kind of knowledge available for use in modeling is not easy. On the one hand, the manual creation of wide-coverage knowledge bases is infeasible, due to the size and complexity of relevant script knowledge. On the other hand, texts typically refer only to certain steps in a script and leave a large part of this knowledge implicit, relying on the reader's ability to infer the full script in detail. Thus, extraction of script knowledge from large text corpora (as done by Chambers and Jurafsky (2009)) is difficult and the outcome can be noisy.

In this work, we present a large-scale crowdsourced collection and annotation of explicit linguistic descriptions of event patterns, to be used for the automatic acquisition of high-quality script knowledge. This work is part of a larger research effort where we seek to provide a solid empirical basis for high-quality script modeling by inducing script structure from crowdsourced descriptions of typical events, and to investigate methods of text-to-script mapping, using naturalistic texts from crowdsourced stories, which describe real-life experiences and instantiate the same scripts (Modi et al., 2016). Predecessors of our work are the OMICS and SMILE corpora (Singh et al., 2002; Regneri et al., 2010), containing multiple event-sequence descriptions (ESDs) for specific activity types or *scenarios*.

Figure 1 shows some example ESDs for the BAKING A CAKE scenario. As can be seen from the examples, the linguistic descriptions of the same event are different, but semantically similar (e.g. mixing: *stir together and mix*, *mix in the extra ingredients*, *mix well*). Also, semantically similar event descriptions tend to occur in relatively similar positions in the ESDs.

1. Take out box of cake mix from shelf
2. Gather together cake ingredients
3. Get mixing bowl
4. Get mixing tool or spoon or fork
5. Add ingredients to bowl
6. Stir together and mix
7. Use fork to breakup clumps
8. Preheat oven
9. Spray pan with non stick or grease
10. Pour cake mix into pan
11. Put pan into oven
12. Set timer on oven
13. Bake cake
14. Remove cake pan when timer goes off
15. Stick tooth pick into cake to see if done
16. Let cake pan cool then remove cake

1. Get a cake mix
2. Mix in the extra ingredients
3. Prepare the cake pan
4. Preheat the oven
5. Put the mix in the pans
6. Put the cake batter in the oven
7. Take it out of the oven

1. Purchase cake mix
2. Preheat oven
3. Grease pan
4. Open mix and add ingredients
5. Mix well
6. Pour into prepared pan
7. Bake cake for required time
8. Remove cake from oven and cool
9. Turn cake out onto cake plate
10. Apply icing or glaze

Figure 1: Example Event Sequence Descriptions (ESDs).

The extraction of structured script information from these descriptions can be viewed as the task of grouping event descriptions into paraphrase sets exploiting semantic and positional similarities, then inducing the script structure from the paraphrase sets. Figure 2 shows a possible induced script structure for the BAKING A CAKE scenario, with nodes representing events in the scenario linked to paraphrase sets of semantically similar linguistic descriptions of the same event.

Regneri et al. (2010) used Multiple Sequence Alignment (MSA, Durbin et al. (1998)) to induce script structure by aligning semantically similar event descriptions across different ESDs. Roughly speaking, the resulting paraphrase sets correspond to the script's event types, while their default temporal order is induced from the order of the event descriptions in the ESDs. The precision of MSA-based script extraction is impressive, but MSA has a fundamental drawback, that is, the strict assumption of a fixed and invariable order for events.

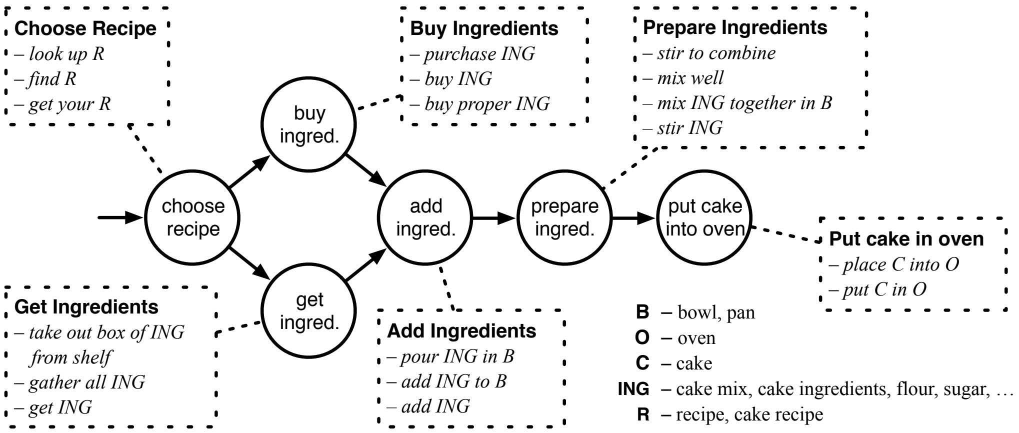


Figure 2: Example of an induced script structure for the BAKING A CAKE scenario.

<u>choose_recipe</u>
Review desired recipe, Look up a cake recipe,
Print out or write down the recipe, Read recipe, ...

<u>buy_ingredients</u>
Buy other ingredients if you do not have at home,
Buy cake ingredients, Purchase ingredients, ...

<u>get_ingredients</u>
Gather all ingredients, Set out ingredients, Gather ingredients,
gather together cake ingredients such as eggs, butter, ...

<u>add_ingredients</u>
Add water, sugar, beaten egg and salt one by one,
*Whisk after each addition,*Add the dry mixture to the wet mixture,
Mix the dry ingredients in one bowl (flour, baking soda, salt, etc),
Add ingredients in mixing bowl, *get mixing bowl, ...*

<u>prepare_ingredients</u>
Mix them together, Open the ingredients, Stir ingredients,
Combine and mix all the ingredients as the recipe delegates,
Mix ingredients with mixer, ...

<u>put_cake_oven</u>
Put the mix in the pans,Put the cake batter in the oven,
Put cake in oven, Put greased pan into preheated stove,
Store any leftovers in the fridge, Cover it and put it on a oven plate
Put the prepared oven plate inside oven ...,

Figure 3: Example clusters (event labels are underlined, outliers are in italics).

Script events are temporally ordered by default, but their order can vary. For example, when baking a cake, one can *preheat the oven* before or after *mixing the ingredients*. MSA does not allow for crossing alignments, and thus is not able to model order variation: this leads to an inappropriately fine-grained inventory of event types.

Clustering algorithms using both semantic similarity and positional similarity information provide an alternative way to induce scripts. They can be made sensitive to ordering information, but do not use it as a hard constraint, and therefore allow for an appropriate treatment of order variation. We have experimented with several clustering algorithms. Figure 3 shows an example output of the best-performing algorithm, Affinity Propagation (AP, Frey and Dueck (2007)). The outcome is quite noisy. For instance, in the last cluster, which mainly consists of descriptions of the "putting cake in oven" event, *put the mix in the pans* and *store any leftovers in the fridge* clearly are outliers. The noise is to some degree due to the complex nature of script structures in general, but it is also the price one has to pay for a gain in flexibility of event ordering.

Clustering algorithms rely on good estimates of similarity among the data points. To appropriately group event descriptions into paraphrase sets, the clustering algorithm would need information on script-specific semantic similarity that goes beyond pure semantic similarity. For instance, in the FLYING IN AN AIRPLANE scenario, it is not trivial for any semantical similarity measure to predict that *walk up the ramp* and *board plane* are functionally similar with respect to the given scenario. To address this issue, we collect partial alignment information that we will use as seed data in future work on semi-supervised clustering. The alignment annotations are also suitable for a semi-supervised extension of the event-ordering model of Frermann et al. (2014).

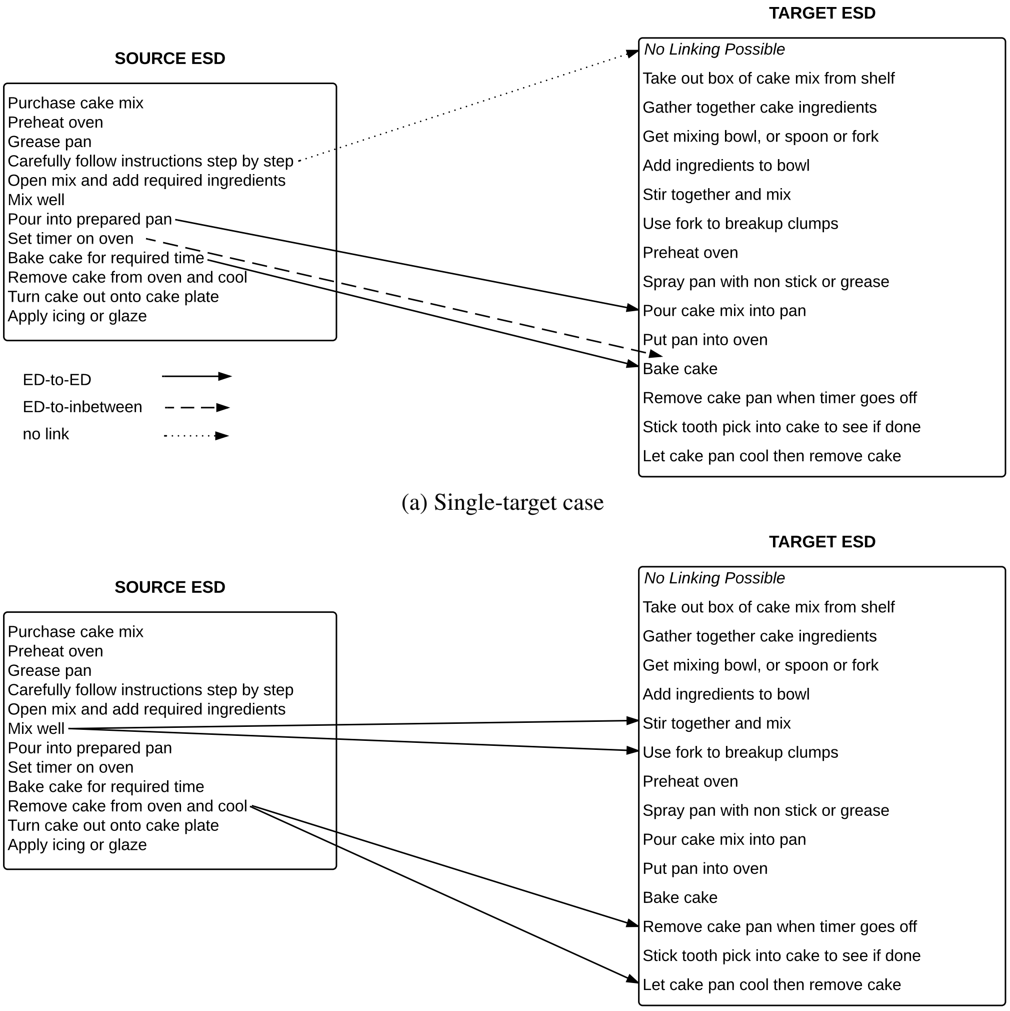
In this work, we have taken measures to provide a sound empirical basis for better-quality script models, by extending existing corpora in two different ways. First, we crowdsourced a corpus of 40 scenarios with 100 ESDs each, thus going beyond the size of previous script collections. Second, we enriched the corpus with partial alignments of ESDs, done by human annotators. The result is a corpus of partially-aligned generic activity descriptions, the **DeScript** corpus (**De**scribing **Script** Structure). More generally, DeScript is a valuable resource for any task involving alignment and paraphrase detection of events.

The corpus is publicly available for scientific research purposes at this url: `http://www.sfb1102.uni-saarland.de/?page_id=2582`.

2. Data Collection

2.1. ESD Collection

Scenario choices were based on previous work by Raisig et al. (2009), Regneri et al. (2010) and Singh et al. (2002). We included scenarios that require simple general knowledge (e.g. WASHING ONE'S HAIR), as well as more complex ones (e.g. BORROWING A BOOK FROM THE LIBRARY), scenarios showing a considerable degree of variablity (e.g. GOING TO A FUNERAL) and scenarios requiring some amount of expert knowledge (e.g. RENOVATING A ROOM).


SOURCE ESD
Purchase cake mix
Preheat oven
Grease pan
Carefully follow instructions step by step
Open mix and add required ingredients
Mix well
Pour into prepared pan
Set timer on oven
Bake cake for required time
Remove cake from oven and cool
Turn cake out onto cake plate
Apply icing or glaze
ED-to-ED
ED-to-inbetween
no link
TARGET ESD
No Linking Possible
Take out box of cake mix from shelf
Gather together cake ingredients
Get mixing bowl, or spoon or fork
Add ingredients to bowl
Stir together and mix
Use fork to breakup clumps
Preheat oven
Spray pan with non stick or grease
Pour cake mix into pan
Put pan into oven
Bake cake
Remove cake pan when timer goes off
Stick tooth pick into cake to see if done
Let cake pan cool then remove cake

(a) Single-target case

(b) Multiple-target case

Figure 4: Examples of possible annotations for the BAKING A CAKE scenario.

Activity descriptions were crowdsourced via Amazon Mechanical Turk (M-Turk)[1]. A total of 320 workers (native speakers of English) described everyday activities in small sequences of short sentences. Each worker could write at most one ESD per scenario, and 5 to 16 event descriptions per ESD. They were paid 0.20 USD per ESD and took on average 2.78 minutes per ESD. After a pilot study on 10 scenarios with 10 ESDs per scenario, we collected the full corpus of 40 scenarios with 100 ESDs per scenario. Once the data was collected, it was manually checked to remove ESDs that had unclear language or where the worker misunderstood the task (7% of the ESDs).

Although the collected dataset has high conformity, there are interesting scenario-specific differences, which can be captured by different metrics. Easier scenarios show lower vocabulary variance than more complex ones as measured by type-token ratio (TTR) per scenario: e.g. TAKING A SHOWER has the smallest vocabulary variance with a TTR of 0.07 and shortest event descriptions on average at 3.89 tokens, while RENOVATING A ROOM has the highest TTR of 0.16 and among the longest event descriptions on average at 4.90 tokens, with the longest being SENDING FOOD BACK (IN RESTAURANT) at 5.52 tokens. ESDs for one scenario or another also differ with regard to the amount of knowledge workers share about them. For example, ESDs for TAKING A SHOWER and RENOVATING A ROOM differ not only for their TTR and length, but also in their homogeneity: ESDs for TAKING A SHOWER are most similar to one another (average Dice ESD word-type overlap of 0.46) while ESDs for RENOVATING A ROOM are least similar to one another (average Dice ESD word-type overlap of 0.2). While everyone has common knowledge about *taking a shower*, *renovating rooms* is not something we all share expertise about or do in the same way.

2.2. Alignment Annotation

A second step in the data collection enriched the ESD corpus with partial alignment information, to be used as seed data in semi-supervised clustering of event descriptions into semantically similar paraphrase sets. The partial alignment information was also crowdsourced via M-Turk. We chose a representative set of 10 scenarios, with approximately 100 ESDs each, to be used in the alignment study. The workers were presented with a source and a target ESD and asked to link highlighted descriptions from the source ESD with those event descriptions from the target ESD that were semantically similar to those in the source ESD (see Figure 4). They had the option of either finding a **single-target** description in the target ESD or **multiple-target** descriptions.

Single target

In the simplest case, workers linked one event description in the source ESD to one event description in the target

[1] https://www.mturk.com/mturk/

ESD (**ED-to-ED link**, e.g. *pour into prepared pan* → *pour cake mix into pan* in Figure 4a). If the target ESD did not contain any matching event description, the workers could either select a position between two event descriptions on the target side where the source event would usually take place (**ED-to-in-between links**, e.g., *set timer on oven* could take place between *put pan into oven* and *bake cake*), or they could indicate that no linking at all is possible (**no-links**). This latter option is useful in case of spurious events (e.g., *carefully follow instructions step by step* is not really an event) but also for alternative realizations of a script (e.g. *paying cash* vs. *with a credit card*).

Multiple target
If the workers felt that the source event was described in more detail in the target ESD compared to the source ESD, they could link the source event description to more than one event description in the target ESD (e.g. *mix well* → *stir together and mix, use fork to break up clumps*, which can be broken down to two or more **ED-to-ED links**, see Figure 4b). Also when linking the source description to multiple descriptions in the target ESD, workers could choose in-between positions (**ED-to-in-between links**).

Seed data selection
In order to minimize the amount of seed data that would be needed for semi-supervised clustering, we employed several criteria for choosing the most informative event descriptions to be aligned. Informative seeds are expected to have the strongest corrective effect on the clustering algorithm. Thus, the event descriptions to be aligned should be the borderline cases that are the most difficult for the algorithm. In order to select the borderline cases, henceforth called *outliers*, we used two methods. First, we ran the Affinity Propagation clustering algorithm varying the preference parameter that determines the cost of creating clusters, thus leading to different configurations of clusters (i.e. to a varying number of clusters and cluster sizes), and we chose those event descriptions that changed their neighbors and cluster centers as the number of clusters increased or decreased. Secondly, we chose those event descriptions that were not well clustered as measured by the Silhouette index, which takes into account the average dissimilarity of an item to the members of its cluster and its lowest average dissimilarity to the members of the other clusters. The two criteria used would select the most difficult cases for the algorithm, that is, cases whose true alignment is expected to be most informative.

We also included event descriptions that were not outliers, henceforth called *stable cases*, that were used as a baseline when evaluating the inter-annotator agreement to show the difficulty of the outliers. The stable cases were randomly selected from those event descriptions that were not outliers. We expect more annotator agreement in stable cases as compared to the outliers. Approximately 20% of the event descriptions per scenario were selected as outliers and 10% were selected as stable cases and presented to the workers to align with another event[2].

Additionally, we included *gold seeds*, that is a small subset of the event descriptions aligned by experts as part of our gold standard annotation (see Section 2.3.), to be used for an evaluation of the workers' annotation. 5% of the event descriptions per scenario were included as gold seeds. The workers were not aware of what alignments were outliers, stable cases or gold seeds.

Each source ESD containing outlier, stable or gold event descriptions was matched with 3 target ESDs. Each source-target ESD pair was annotated by 3 different annotators. In total approximately 600 outlier event descriptions, 300 stable event descriptions and 120 gold seeds were selected for each scenario and presented to workers for annotation. 292 workers (native speakers of English) took part in the annotation study. They were paid 0.35 USD per ESD and took on average 1.05 minutes per ESD.

2.3. Gold Standard Alignment Annotation

The DeScript corpus also includes a gold standard corpus of aligned event descriptions annotated by experts, to be used in the evaluation of semi-supervised clustering of event descriptions. A small subset of the gold standard alignments were used to evaluate the quality of the crowd-sourced alignments. The gold standard covers 10 scenarios, with 50 ESDs each.

Four experts, all computational linguistics students, were trained for the task and were presented with a source and a target ESD in an interface where (unlike the M-Turk interface) no event description in the source ESD was highlighted. They were to link all event descriptions in the source ESD with all event descriptions in the target ESD that were semantically similar, with respect to the given scenario, to those in the source ESD. Every ESD in a given scenario was paired with every other ESD in the same scenario. In the end, all similar event descriptions were aligned and the full alignments were used to group the event descriptions into gold paraphrase sets.

Scenario	**EDs**		**Gold**
	annotated	**excluded**	**sets**
baking a cake	513	29	20
borrowing a book from the library	389	46	16
flying in an airplane	504	46	24
getting a haircut	418	52	23
going grocery shopping	505	95	18
going on a train	357	45	12
planting a tree	344	41	13
repairing a flat bicycle tire	363	40	16
riding on a bus	358	23	14
taking a bath	479	29	20
Total/Average	4230	446 (10.5%)	18

Table 1: Gold alignment annotation: the annotated EDs for each scenario, the excluded EDs for each scenario (singletons or unrelated events) and the number of gold paraphrase sets obtained.

[2] Note that we refer to the number of descriptions per scenario. The annotated alignments are less than the 1% of all possible alignments (c.a. 0.1-0.2%).

Scenario	**ED-to-ED links**	**ED-to-in-between links**	**No-links**	**Total links**
baking a cake	1836	831	448	3115
borrowing a book from the library	1787	1146	254	3187
flying in an airplane	1676	1219	192	3087
getting a haircut	1887	926	292	3105
going grocery shopping	1863	997	251	3111
going on a train	1937	884	285	3106
planting a tree	1902	845	343	3090
repairing a flat bicycle tire	1500	1019	612	3131
riding on a bus	2226	750	114	3090
taking a bath	1898	891	314	3103
Total	18512	9508	3105	31125

Table 2: All links drawn for all source-target ESD pairs by all annotators. Note: the first column includes both ED-to-ED links from *single-target* cases and each single ED-to-ED link (arrow) in *multiple-target* cases.

In creating the gold paraphrase sets, we assumed script-specific functional equivalence, that is, we instructed the annotators to group event descriptions serving the same function in the script as semantically similar (e.g, *scan bus pass*, *pay for fare* and *show driver ticket* are grouped into the same paraphrase set in the RIDING ON A BUS scenario, as they all represent the "pay" event). Each paraphrase set was annotated with an event label that indicated the event being expressed in the given paraphrase set (e.g. in Figure 2, *choose_recipe* and *buy_ingredients* are example event labels for BAKING A CAKE). Event labels were harmonized with those used in the annotation of stories in Modi et al. (2016) (see Section 4.).

Paraphrase sets that were singletons (e.g. in BAKING A CAKE, *return to oven* occurred only once) were considered not representative of the events in the scenario, and hence were removed based on the gold annotation. Likewise, event descriptions that were not related to the script (e.g., in RIDING ON A BUS, *pray your bus is on time*) or that were related to the scenario but not really part of the script (e.g., in BAKING A CAKE, *store any leftovers in the fridge*), were also removed. In the end, approximately 10.5% of the annotated event descriptions were excluded. Table 1 indicates the number of excluded event descriptions and the number of gold clusters per scenario.

The result of the gold standard annotation is a rich resource of full alignments for all event descriptions in 500 ESDs (10 scenarios with 50 ESDs each) grouped in gold paraphrase sets, each containing different linguistic variations of the events in the scenario.

3. Data Analysis and Evaluation

The alignment links we collected (Table 2) show an interesting degree of variability across scenarios and across the ESDs within a scenario. A high number of *ED-to-in-between links* and *no-links* shows that not all events are verbalized in every ESD for the same scenario: the source ESD may contain optional events which are either not explicitly mentioned in the target ESD (ED-to-in-between links) or event descriptions in source ESD that are not considered events in the scenario (no-links). Recall that workers could either find a single-target description in the target ESD or multiple-target description. They showed a strong preference for single-target links (which were chosen 30021 times) over multiple-target links (which were chosen 238 times).

We distinguish between *one-to-one* alignments, that is, cases which all three annotators considered to be *single-target* cases, since they used exactly one link between source and target ESD (including *in-between-links* and *no-links*), and *one-to-many* alignments, that is cases which at least one annotator considered to be *multiple-target* cases, using more than one link between source and target ESD. Note that, as annotators preferred single-target links over multiple-target links, most source event descriptions are annotated as *one-to-one* alignments (Table 3).

Many workers were involved in the alignment annotation task and not all of them aligned the complete set of highlighted event descriptions. For this reason, we computed agreement as the number of times where the majority of workers agreed in an alignment instance, instead of the Kappa (Fleiss, 1971).

One-to-one alignments

Figure 5 shows how well the workers agreed with each other in the annotation of the *outliers* and *stable cases* for *one-to-one* alignments. As expected, on average, workers tended to agree more in *stable cases* as compared to *outliers*. The relatively low agreement figures for BORROWING A BOOK FROM THE LIBRARY can be explained by the variety and complexity of the scenario: e.g. given a source event description *find the book on the shelf*, and *select a book* and *take the book off the shelf* on the target ESDs, it is not obvious if the source event is most similar to *find the book on the shelf*, *take the book off the shelf* or *in between* the two.

One-to-many alignments

In the *one-to-many* cases, 79% of event description instances have at least partially overlapping annotations, that is, there is an overlap in the alignments of two or all three annotators (see Table 3). We calculated the average Dice (indicating the degree of overlap between two sets) for all possible pairs of worker annotations for a given *one-to-many* alignment. The BORROWING A BOOK FROM THE LIBRARY scenario has the highest number of *one-to-many* alignments (55) and the highest number of non-overlapping alignments (25), and REPAIRING A FLAT BICYCLE TIRE scenario has the lowest Dice score (0.2). The

Scenario	**One-to-one alignments**			**One-to-many alignments**		
	tot	maj. agreem.	% agreem.	tot	overlap	avg Dice
baking a cake	1091	885	0.81	37	28	0.38
borrowing a book from the library	1081	718	0.66	55	30	0.4
flying in an airplane	1081	904	0.84	14	13	0.36
getting a haircut	1003	810	0.81	13	12	0.44
going grocery shopping	993	856	0.86	28	23	0.41
going on a train	982	810	0.82	34	32	0.45
planting a tree	1000	822	0.82	21	18	0.63
repairing a flat bicycle tire	991	708	0.71	29	17	0.2
riding on a bus	1074	914	0.85	28	25	0.46
taking a bath	1101	933	0.85	26	24	0.38
Total	10397	8360	0.80	285	222	0.41

Table 3: Number of source event descriptions that all workers annotated as single-target (*one-to-one* alignments), and number of descriptions where at least one chose the multiple-target option (*one-to-many* alignments), with majority counts and overlap counts.

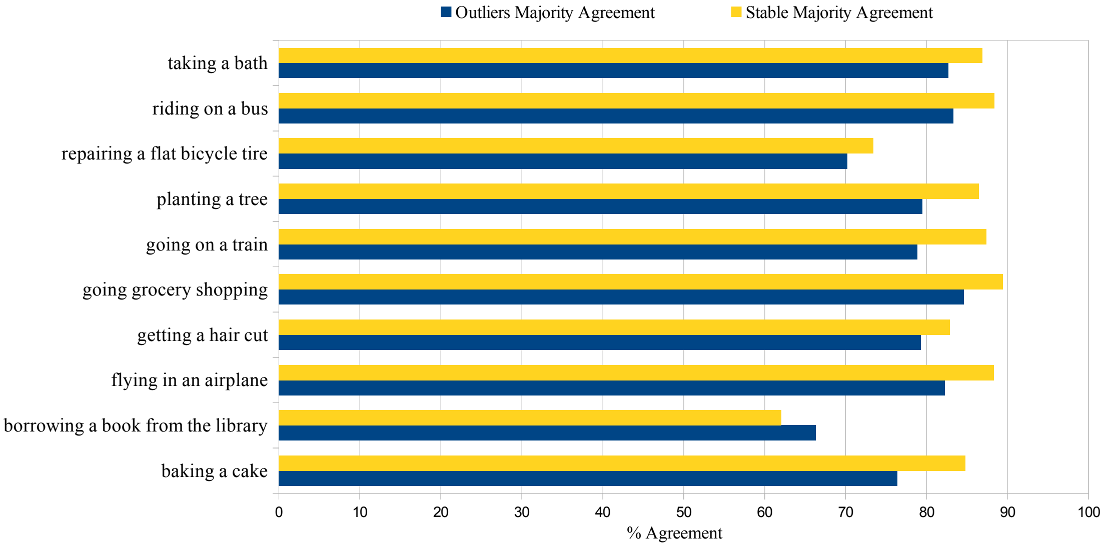


Figure 5: Worker agreement for *outliers* and *stable cases* in *one-to-one* alignments.

two scenarios are among those with the highest complexity and variability in how the script could be carried out.

The overall average Dice is 0.41, that is, the agreement is quite good on corresponding core events, although there may be disagreement about the precise sequence in the target ESD that corresponds to a given source event. For instance, in Figure 4b workers may agree on the corresponding core event (as *mix well* is semantically similar to *stir together and mix*), but they may not agree on the corresponding span, whether it is most similar only to *stir together and mix* or it also entails *use fork to break up clumps*.

Gold seeds

Recall that besides the outliers (the difficult cases) and the stable cases (which were used as a baseline), we also included gold seeds for evaluation purposes, in order to compare the workers' annotation against expert annotation. Unsurprisingly, majority agreement in both *stable cases* and *gold seeds* was higher than agreement on *outliers* (87% for *gold seeds*, 82% for *stable cases* and 78% for *outliers*), showing that, while our outlier selection method effectively selects more challenging cases, the quality of the annotation is still very satisfactory.

Agreement between the worker's majority vote and the gold annotation was 81%. The cases where the workers did not agree with the gold annotation also illustrate the inherent complexity of the scripts. For example, in BORROWING A BOOK FROM THE LIBRARY, the workers aligned *take the book home* in the source ESD to the position between *leave library* and *read book* in the target ESD, while the experts aligned the same description to *leave library*. This shows that the workers were not necessarily wrong in all the cases where they did not agree with the *gold seeds*.

No-links

Interestingly, the workers tended to use *no-links* for those event descriptions that were not really events (e.g., in BAKING A CAKE: *carefully follow instructions step by step*) or event descriptions that were unrelated to the given scenario (e.g., in RIDING ON A BUS: *sing if desired*). The event descriptions annotated as *no-links* by the workers tend to overlap with those marked by experts as unrelated event descriptions in the gold standard. These cases typically involve event descriptions that are misleading and should not be part of the script. This shows that certain event descriptions are spurious, cases which we cannot expect a clustering algorithm to group in any meaningful way.

4. Comparison with the InScript Stories

As mentioned in the introduction, this work is part of a larger research effort where we seek to provide a solid empirical basis for high-quality script modeling. As part of this larger project, Modi et al. (2016) crowdsourced a corpus of simple scenario-related stories, the InScript corpus (Narrative Texts **In**stantiating **Script** structure). They asked workers on M-Turk to write down stories narrating recent real-life experiences instantiating specific scenarios (e.g. EATING IN A RESTAURANT). The induced script structure from the ESDs will be used to investigate methods of text-to-script mapping, as well as to model the instantiation of script structures in naturalistic texts from the crowdsourced stories, as depicted in Figure 6.

Modi et al. (2016) created *script templates* that described script-specific event labels and participant labels for each scenario (e.g., event labels in GOING TO A RESTAURANT: *get_restaurant, take_seat, look_menu* and participant labels: *restaurant, waiter, menu*), which were used to annotate the stories. They annotated event-denoting verbs in the stories with the event labels and participant-denoting NPs with the participant labels. Event labels used in the annotation of stories in the InScript corpus were harmonized with the gold paraphrase sets from the DeScript corpus (Section 2.3.) to reach a one-to-one correspondence.

We compared the two resources with regard to their lexical variety, which is higher in the narrative texts than in the ESDs. We chose not to use the type-token ratio (TTR), as it is known to be sensitive to text length, and in this case the narrative texts are generally longer and would result in very low TTR values for the InScript data. Instead, we compared the lexical variety using the Measure of Textual Lexical Diversity (MLTD, McCarthy and Jarvis (2010)), which computes the mean length of word strings that are needed to maintain a set threshold level of lexical variation. We used a threshold TTR value of 0.71, which was empirically set by the authors (high MTLD corresponds to high lexical variety). We noted that the narrative texts have higher MTLDs across all scenarios, ranging between 40 and 47, as compared to ESDs with MTLDs ranging between 26 and 44 (Figure 7). That is, in the narrative texts more tokens are needed to reach the set TTR of 0.71; hence, the narrative texts are more lexically diverse in comparison to the ESDs.

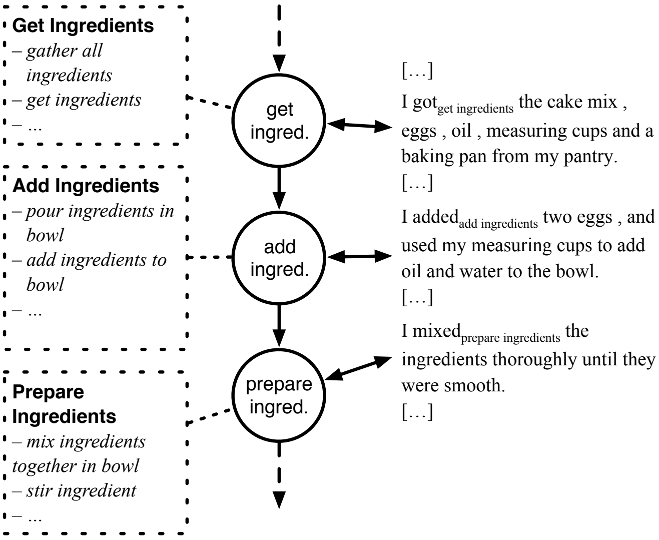

Get Ingredients
– gather all ingredients
– get ingredients
– ...
Add Ingredients
– pour ingredients in bowl
– add ingredients to bowl
– ...
Prepare Ingredients
– mix ingredients together in bowl
– stir ingredient
– ...
get ingred.
add ingred.
prepare ingred.
[...]
I got$_{\text{get ingredients}}$ the cake mix , eggs , oil , measuring cups and a baking pan from my pantry.
[...]
I added$_{\text{add ingredients}}$ two eggs , and used my measuring cups to add oil and water to the bowl.
[...]
I mixed$_{\text{prepare ingredients}}$ the ingredients thoroughly until they were smooth.
[...]

Figure 6: Connecting DeScript and InScript: an example from the BAKING A CAKE scenario (InScript participant annotation is omitted for better readability).

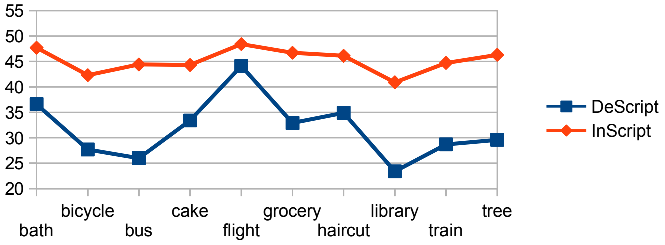


Figure 7: MTLD values for DeScript and InScript, per scenario.

As expected, the DeScript corpus, a collection of generic descriptions of script-related activities, has a lower lexical diversity compared to the InScript corpus, which in turn contains naturalistic texts describing real-life experiences. Our next goal is to exploit script structures induced from the ESDs in DeScript for a text-to-script mapping of the script-related naturalistic texts in InScript.

5. Conclusions

We collected a corpus of 3,948 event sequence descriptions (40 different scenarios, approximately 100 different event sequence descriptions descriptions per scenario), ranging from simpler ones to ones that show interesting variation with regard to their granularity, to the events described, and to different verbalizations of the same event within a scenario. The corpus, which is to our knowledge the largest collection of event sequence descriptions available, is enriched with partial alignment information on difficult event descriptions. *Multiple-target* annotations and *in-between* links are of particular interest, because they can capture differences between event descriptions in terms of granularity and optionality of events.

We also collected full alignments by experts for 10 different scenarios (50 event sequence descriptions per scenario), grouped into labeled paraphrase sets, to be used in the evaluation of semi-supervised clustering of event descriptions. We expect that the crowdsourced corpus and the gold standard alignment set will provide a sound basis for high-quality script models and will be used as a valuable resource for any task involving alignment and paraphrases of events.

6. Acknowledgements

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

7. Bibliographical References

Barr, A. and Feigenbaum, E. A. (1981). Frames and scripts. In *The Handbook of Artificial Intelligence*, volume 3, pages 216–222. Addison-Wesley, California.

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 602–610, Suntec, Singapore.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Frermann, L., Titov, I., and Pinkal, M. (2014). A hierarchical Bayesian model for unsupervised induction of script knowledge. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–57, Gothenburg, Sweden.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science Direct*, 315(5814):972–976.

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*, 42(2):381–392.

Modi, A., Anikina, T., Ostermann, S., and Pinkal, M. (2016). InScript: Narrative texts annotated with script information. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 16)*, Portorož, Slovenia.

Raisig, S., Welke, T., Hagendorf, H., and der Meer, E. V. (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*, pages 979–988, Uppsala, Sweden.

Schank, R. C. and Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ: Erlbaum.

Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., and Li Zhu, W. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In Robert Meersman et al., editors, *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237. Springer, Berlin / Heidelberg, Germany.