Schema Learning Corpus: Data and Annotation Focused on Complex Events

Song Chen, Jennifer Tracey, Ann Bies, Stephanie Strassel
Linguistic Data Consortium
3600 Market Street, Suite 810, Philadelphia, Pennsylvania, United States
{zhiyi, bies, strassel}@ldc.upenn.edu

Abstract

The Schema Learning Corpus (SLC) is a new linguistic resource designed to support research into the structure of complex events in multilingual, multimedia data. The SLC incorporates large volumes of background data in English, Spanish and Russian, and defines 100 complex events (CEs) across 12 domains, with CE profiles containing information about the typical steps and substeps and expected event categories for the CE. Multiple documents are labeled for each CE, with pointers to evidence in the document for each CE step, plus labeled events and relations along with their arguments across a large tag set. The SLC was designed to support development and evaluation of technology capable of understanding and reasoning about complex real-world events in multimedia, multilingual data streams in order to provide users with a deeper understanding of the potential relationships among seemingly disparate events and actors, and to allow users to make better predictions about how future events are likely to unfold. The Schema Learning Corpus will be made available to the research community through publication in Linguistic Data Consortium catalog.

Keywords: language resources, complex events, event schemas, event extraction, information extraction

1. Introduction

Many real-world events are not simple occurrences, but complex phenomena that are composed of numerous subsidiary elements, some of which happen simultaneously, while others are sequential and dependent on each other. Humans routinely reason and make predictions about real-world events by utilizing internalized narrative structures, or schemas, that identify the commonly occurring steps, temporal constraints and roles in a complex event (Piaget, 1965). Systems capable of reasoning about real-world events through generalized schemas provide users with a deeper understanding of the potential relationships among seemingly disparate events and actors, and allow users to make better predictions about how events are likely to unfold in the future.

The Schema Learning Corpus (SLC) is a new linguistic resource designed to support research into the structure of complex events in multilingual, multimedia data. The SLC was created to support technology development and evaluation within the DARPA Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS) Program. KAIROS aims to build technology capable of understanding and reasoning about complex real-world events in order to provide actionable insights to end users (DARPA, 2018). KAIROS systems utilize formal event representations in the form of schema libraries that specify the steps, preconditions and constraints for an open set of complex events; schemas are then used in combination with event extraction to characterize and make predictions about real-world events in a large multilingual, multimedia corpus.

The SLC includes a large pool of background data in English, Spanish and Russian, collected from text, image, video and multimedia data sources that cover a wide variety of events in both formal and informal genres. The corpus defined a set of 100 complex events (CEs) covering 12 different information domains. In addition to the background data, the English and Spanish portions of the SLC include an additional set of manually selected documents that, taken as a whole, provide evidence for the individual steps that comprise each of the defined CEs. These documents were manually labeled to provide document provenance for each step of the CE. In addition, events and relations relevant to the CE were also labeled along with their arguments.

2. Related Work

The Schema Learning Corpus makes several new contributions to the linguistic resource landscape, both in terms of the type of data included and the nature of the annotation.

An important aspect of KAIROS research is the goal of developing technologies that can reason about complex real-world events from diverse information streams. The diversity of data in the KAIROS Schema Learning Corpus -- which includes formal and informal news text, amateur image and video data, instructional materials, social media sources and other informal and non-traditional data types -- represents a great challenge for the rapid comprehension of real-world events. Most existing event corpora focus on a single media type and a limited set of genres, especially for Spanish. For example, the Berkeley FrameNet Project (Baker et al., 1998) is a machine-readable database of frames and lexical units extracted from English text documents which describe event and state, including event participant and how event types relates to each other. The ECB (Bejan and Harabagiu, 2010) and ECB+ (Cybulska and Vossen, 2014) corpora have event coreference annotation across English news articles. The ACE corpora (Mitchell et al., 2004; Walker et al., 2006) have event and relation annotations of Arabic, Chinese and English and multi-genre documents, but all documents are either text or audio transcripts. Similarly, the DEFT ERE corpora (Chen et al., 2020; Chen et al., 2023) focuses mainly on events presented in formal and informal text data in Chinese, English, and Spanish. Li et al. (2022) is an event-rich

image-caption dataset with 106,875 images, including extracted event knowledge. The AIDA corpus (Tracey et al., 2022) contains multimedia source data in English, Russian and Ukrainian, with annotation of entities, events, and relations to support extraction and understanding of conflicting information.

What is more, these prior corpora focus on event, relation, and entity annotations in relative isolation, without indicating how they may be components of a larger complex event. Hovy et al. (2013) shows that subevents can form a particular hierarchical event structure and examines a two-stage approach to finding and improving subevent structures. However, that dataset is limited to annotator of 100 English text documents in the Intelligence Community domain. Similarly, the Event Sequencing Dataset (Mitamura et al., 2017) is labeled for event sequences within a script, including AFTER relations between events and SUBEVENT links between a parent and a child events, but it includes only 360 English text documents.

The SLC incorporates both annotation of complex events and representation of those events in multilingual and multimedia data. The corpus includes event, relation and argument annotations with links to each complex event step, with the goal of supporting the development and evaluation of technology capable of generalizing schemas, extracting instances from real-world data, and finding important missing pieces or predicting outcomes in an evolving complex event.

3. SLC Corpus Design

3.1 Background Data

The source data for the Schema Learning Corpus consisted of multimedia (text, image, and/or video) data in English, Spanish, and Russian. English and Spanish were collected during initial construction of the SLC in the first phase of KAIROS, and Russian was added during the second phase of the program. A portion of the SLC documents were directly relevant to specific complex events, while other documents were of unspecified topic content; these unspecified topic documents comprise the background data.

The SLC contains over 16.2 million background documents, including more than 125,000 audio, video, image or multimedia documents. The data includes a combination of Spanish, English and Russian corpora available in the Linguistic Data Consortium catalog, plus new data collected from websites that were rich in appropriate data in order to fill out the data volume and diversity required for the corpus. New data collection focused particularly on resources for Spanish, including instructional documents (e.g., how-to articles), business and logistics domain documents, and multimedia data.

URLs identified in manual data scouting, as well as full site information for the automatic background collection, were fed into our multimedia data collection pipeline, which collected all text, image, and video elements of the specified web page, processed them into separate files, and recorded metadata maintaining the association between the URL and each element of the page, as well as additional information such as publication date, download date, and relative position of each element on the original web page. The data collection pipeline is designed to ensure suitability of data for downstream annotation and evaluation, with the goal that data developed within KAIROS should be broadly shareable outside the program (e.g., in open evaluations, publication of corpora). Table 1 shows the amount of background data for each language.

Language	Text-only doc count	Multimedia doc count	Total doc count
English	11,877,669	100,282	11,977,951
Spanish	3,822,559	14,942	3,837,501
Russian	435,017	12,652	447,669
All Langs	16,135,245	127,876	16,263,121

Table 1: SLC Background Data by Language and Modality.

3.2 Complex Event Data

To ensure that sufficient content to support annotation and schema development was present in the corpus, we seeded the SLC with documents specifically collected for their relevance to a set of pre-defined complex events. Annotators used CE profiles to guide their search for documents on the web, with the goal of locating a diverse set of documents with respect to media types, languages, and representation of different realizations of steps in the CE. For each URL identified, annotators provided information about the language, presence of relevant text, image, and/or video, and presence of answers to specific queries about each event.

The SLC CEs cover a broad range of domains and are not specifically tailored to any particular scenario selected for KAIROS evaluation, although some effort to ensure coverage of evaluation scenario-relevant (or scenario-adjacent) CEs was part of the selection process for CEs.

3.2.1 Development of SLC Complex Events

A CE is a complex event that contains multiple interrelated steps. For instance, *Dining in a Restaurant* is a CE, with steps that occur in a relatively predictable order, for instance:

1. Guests arrive at the restaurant
2. Guests are greeted by a host
3. Guests are seated
4. Guests study the menu
5. Guests place their order for the meal
6. The meal is prepared
7. The prepared meal is delivered to the guests
8. Guests consume the meal
9. Guests pay the bill
10. Guests depart the restaurant

The steps defined for each CE are not intended to describe every possible variation in how things may play out; instead, they describe the typical way the complex event unfolds. Some steps in the CE may be optional -- for instance, guests may be greeted by a host and may seat themselves immediately after arriving at a restaurant, or guests may be very familiar with the menu and place their order right after being seated. Other steps may be ordered differently -- for instance, guests may pay right after placing the order. On the other hand, some steps are strictly ordered; for instance, guests can't consume a meal until after it is delivered.

The SLC defines a set of 100 CEs designed to align with KAIROS research goals. Some are very specific, such as *Shopping in a Store*. Many are more general, such as *Commercial Travel*. CEs cover a wide range of domains, and are intended to have the following qualities:

- They are events in the world that consist of multiple steps, and the steps can be anchored in event mentions with types compatible with the KAIROS program tag set.
- They are general enough to give rise to multiple schemas, and they cover a wide range of domains and topics.
- They are sufficiently well-represented in the KAIROS data and are likely to have coverage across multiple genres and media types.

We started the process of developing CEs by identifying a wide range of domains and topics, including areas relevant to existing KAIROS evaluation scenarios such as conflict, civil unrest, illegal activities, movement, business workings, as well non-evaluation topics such new development of capabilities, medical intervention, government workings, social life and so on. After considering over 100 candidate topical areas, we defined 12 general domains for the SLC that would be used to guide data scouting and document collection. The purpose of defining KAIROS domains was to ensure that the resulting SLC would generalize to many different kinds of complex events. The 12 domains are:

- Business workings
- Civil unrest
- Conflict or threat
- Disaster
- Government workings
- Cyber or information
- Illegal activities
- Legal proceedings
- Medical intervention
- Movement or travel
- New capability development
- Social life

Each domain has 3 or more CEs of various granularities. For example, the Disaster domain has the following CEs:

- Disease outbreak
- Evacuation
- Provide or distribute disaster relief
- Search and rescue
- Supply shortage

CEs in the SLC are descriptive and human-readable, rather than being formal representations and machine readable. We manually created a profile for each CE based on a template to inform data scouting for CE-relevant documents and to provide both annotators and system developers with key information about the CE. CE profiles include a natural language description of the CE along with a set of typical steps that comprise the event. These steps are described in natural language and include information about the expected event tag set types that might instantiate the step, along with information about the expected ordering of the step with respect to other steps. Each CE profile was further formalized to include a unique CE ID, CE steps and step names to support data scouting and annotation. Figure 1 contains the CE profile for the CE *Obtain or Provide Medical Treatment*.

ID: ComplexEvent008
Title: Obtain or Provide Medical Treatment

Description: Medical treatment is applied to one or more people due to a condition or injury. This event can include communications requesting or offering help, transportation to a medical facility, and any interactions with medical personnel or others as part of the process of diagnosing and treating the patient's condition. The focus of this event is on medical treatment by professionals such as doctors, nurses, paramedics, etc., but some treatments may also be applied by non-professionals (e.g., a bystander may perform CPR while waiting for an ambulance to arrive).

Scope of event: The event begins when someone has a medical need and seeks treatment (or someone else seeks treatment on their behalf). The need may arise from illness, injury, accident, or other causes, but the cause itself is not part of the event. The event ends when treatment ceases.

Step 1 RequestTreatment: Medical treatment requested
Expected event types: Communication
Ordering: May be optional or inferred; Should precede Step 2

Step 2 TravelForTreatment: Travel to bring patient and medical personnel together.
Expected event types: Movement (of people and/or things)
Ordering: May be optional or inferred; Should follow Step 1; May precede or concurrent to Step 3
Step 2.1 TransportMedicalPersonnel: Medical personnel travel to the location of the person needing treatment
Expected event types: Movement (of people and/or things)
Step 2.2 TransportPatient: Patient transported to medical facility
Expected event types: Movement

Step 3 Diagnosis: Medical personnel diagnose patient
Expected event types: Inspection (physical), Test (function), Communication
Ordering: May be optional or inferred; May follow or concurrent to Step 2 and precede Step 4

Step 4 Treatment: Medical personnel treat patient
Expected event types: Inspection (physical), Testing (function), medical treatment
Ordering: Required; Should follow Step 3

Figure 1: CE profile for CE *Obtain or Provide Medical Treatment*.

Steps in the CE can have substeps to reflect differing granularity of steps. Substeps are either more specific steps (but more coarse-grained than atomic events such as contact, movement, transaction) that can happen in a sequence or more specific alternatives for the step. For instance, in the CE *Non-Violent Protest*, one of the steps is "Plan the Protest", that might include substeps "Secure Permit", "Get Funds", "Publicize" which are finer-grained steps than "Plan the Protest". Alternatively, in the CE *Obtain or Provide Medical Treatment*, the step "Travel to bring patient and medical personnel together" has two alternatives as substeps, either "Medical personnel travel to the location of the person needing treatment" or "Patient

transported to medical facility". The ordering information represents common or expected ordering rather than attempting to encode every possible edge case.

There could be multiple ways to define steps or substeps in a complex event. In our case, they are identified manually by trained human annotators following the basic framework established for the KAIROS program at its outset by DARPA program manager, which includes validating that the defined steps or substeps do appear in the corpus via manually scouting for documents that contain evidence that instantiates the steps or substeps.

3.2.2 Data Scouting

The CE profiles serve as a guide to scouting and annotation in the SLC. During data scouting, annotators consult the CE profile, and search the web not just for documents that discuss that CE, but also for documents that have evidence for the specific steps involved in the CE. A key part of data scouting is ensuring that we have sufficient variety in the data collected for each CE and its steps, in terms of the data source, genre, modality and language of the documents. We also ask data scouts to find highly varied examples of the step evidence itself, so that the details for each step -- the who, what, when, where and so on vary across the collected instances. Information about the collected data and the features for each step is tracked in a central database so that annotators as a group can easily see where more variety is needed and adjust their scouting strategies accordingly. For example, data scouting for the CE *Provide and Distribute Disaster Relief* was assigned to three English and three Spanish annotators, each looking for documents on this CE in their language. Table 2 shows some real-world events in documents that annotators scouted from the web which have events or relations that instantiate the steps of the CE.

Lang	Real-world Event	Modality
Eng	A group of people raise money for Hurricane Sandy	Text
Eng	Collecting money to help flood victims in Kerala, India	Multi-media
Eng	Supplies shipped to Bahamas after Hurricane Dorian	Video
Spa	Red Cross' response to 2017 earthquake in Mexico City	Video
Spa	Humanitarian aid arrives in Venezuela	Multi-media

Table 2: Real-world events in documents scouted for CE *Provide and Distribute Disaster Relief.*

Annotators record the CE steps they find in the documents during scouting and also note the media type for the element in the document that instantiates a given CE step. This information is collated in a table (as shown in Figure 2) and presented to all annotators working on the same CE in real time, so that they can see at a glance which steps require more scouting to support the goal of finding documents representing all of the steps across multiple data types.

CE Step Number	CE Step Name	Docs with Text	Docs with Audio	Docs with Image	Docs with Video
1	Collect Aid	4	1	1	2
1.1	Appeal for Aid	6	2	1	1
1.2	Collect Physical Aid	4	1	2	3
1.3	Collect Financial Aid	4	3	1	1
1.4	Purchase Aid Items	1	1	0	1
2	Transfer Aid to Region	3	1	0	2
2.1	Transport Aid to Region	3	1	1	2
2.2	Transfer Funds	3	2	1	1
2.3	Distribute Aid	5	1	1	3

Figure 2: Tracking scouting yield for CE *Provide and Distribute Disaster Relief* for each CE step.

In total, we scouted and collected over 3,600 multilingual documents for 100 CEs, with 2/3 being English and 1/3 being Spanish. Note that no CE-relevant documents were manually scouted or annotated for Russian because this language was added to the KAIROS program after the original SLC was complete. However, Russian background data was added to the SLC as discussed in section 3.1 above.

Table 3 shows the number of documents newly scouted and collected for each domain in each language.

Domain	CEs	Eng docs	Spa docs	Total docs
Business Workings	16	395	219	614
Civil Unrest	3	65	36	101
Conflict Threat	14	248	152	400
Cyber Information	7	127	88	215
Disaster	5	116	65	181
Government Workings	10	272	128	400
Illegal Activity	7	150	97	247
Legal Proceedings	9	290	138	428
Medical Intervention	4	103	36	139
Movement	3	104	55	159
New Capacity Development	16	339	177	516
Social Life	6	189	102	291
Total	100	2398	1293	3691

Table 3: Scouted and collected documents for each domain in each language.

3.3 Annotation

Annotation of CEs in the SLC focuses on providing examples of event and relation mentions that make up the steps of a CE rather than exhaustive annotation of all mentions in the data. The events we defined and labeled in the SLC are intended to serve as exemplars to illustrate both the kinds of complex events and also the steps or substeps for those complex events that are present in the data. Documents were selected for annotation based on

information recorded during scouting, balancing the coverage of CE steps and variety of language, media and genre. Each selected document was annotated with respect to only one CE; any event mentions in the document that may have been related to a different CE were disregarded for annotation purposes.

3.3.1 Annotation Tag Set

The tag set used to label entity, event, and relation types is the same as the KAIROS Phase 1 annotation tag set, which evolved significantly over the course of the KAIROS program and represented a significant expansion over annotation tag sets utilized in prior DARPA programs like DEFT (Song et al., 2015). The motivation for this expansion was two-fold. First, it was intended to cover CE-relevant areas of information that were absent from the prior annotation tag sets (including cognitive events such as research). Second, it was intended to provide finer-grained subtypes and sub-subtypes that KAIROS systems might be able to utilize in developing event schemas. For example, rather than having a single Movement.Transportation event type, the SLC tag set breaks this into 5 different categories, distinguishing evacuation and illegal transportation from other kinds of transportation movement. Every event and relation type in the SLC also included an "unspecified" subcategory (for instance, Movement.Transportation. Unspecified) that annotators could use as a backoff category when the more detailed subtype of the event/relation was not clear from the data, or when the relevant detailed sub-subtype was not present in the tag set.

Table 4 presents information about the SLC annotation tag set.

Category	**SLC Tag Count**
Event types	67
Event types with event arguments	31
Relation types	46
Relation types with event arguments	9
Entity/filler types	24
Total types	137

Table 4: Size of KAIROS SLC Annotation Tag Set.

3.3.2 Provenance Linking Annotation

The first type of annotation performed on SLC documents was Provenance Linking, a lightweight approach to grounding the presence of CE steps in documents. This approach was adopted to provide a first layer of annotation that emphasized the linking of events in documents to steps in a CE (using the CE profile as a stand-in for a schema), which is the primary focus of the KAIROS program. For instance, a multi-media English document was collected for the CE *Provide And Distribute Disaster Relief*. In Example 1, annotators found evidence for the step "Collect Physical Aid" in the video element as seen in Figure 3, and marked a specific timespan in the video as the provenance. They also assigned an event type from the tag set (in this example, Transaction.Donation.Unpsecified) to this event instance. CE steps may be instantiated across different documents, languages and modalities. In Example 2, evidence for another step for this CE, "Transport Aid to the Region", comes from an entirely different document, modality and language; and it is about a different disaster - Hurricane Dorian in the Bahamas rather than flooding in the UK.

Example 1: Video element from English document on Doncaster Flooding, England, Nov 2019

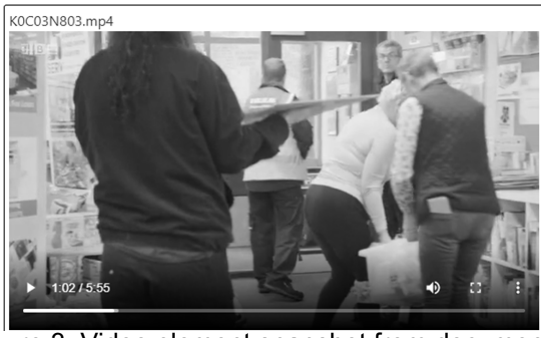

Figure 3: Video element snapshot from document on Doncaster flooding, England, Nov 2019.

- Event Type: Transaction.Donation.Unspecified
- Linked step: Step 1.2: Collect Physical Aid
- Provenance: video timestamp 0:58-1:05

Example 2: Text element from Spanish document on Hurricane Dorian, Bahamas:

USAID transportó vía aérea 47 toneladas métricas de suministros de socorro crítico el 4 de septiembre desde su almacén de emergencia en Miami a las Bahama.

English Gloss: USAID airlifted 47 metric tons of critical relief supplies Sept. 4 from its emergency warehouse in Miami to the Bahamas.

- Event Type: Movement.Transportation.Unspecified
- Linked step: Step 2.1: Transport Aid to Region
- Provenance: text offset range 2229-2759

3.3.3 Event and Relation Mention Annotation

Event and relation Mention Annotation in the SLC provides more detailed and structured representation in the form of event and relation frames for event and relation mentions found in the same documents that were previously annotated for provenance linking. Mention Annotation fleshes out the approximate provenance spans in the Provenance Linking task to a carefully selected span for each event and relation trigger, while the basic event or relation type from Provenance Linking is reviewed and updated if needed, using the three-level typing (type, subtype and sub-subtype) from the KAIROS annotation tag set. In addition, the frame includes the mention attribute, argument roles, and entity mentions that fill

the argument slots, the spans for those arguments, start and end timestamp (for event mentions only), and importantly, which step of the CE the mention is linked to. By linking labeled events and document provenance to steps or substeps in the CE profile, the schema structure was instantiated with real occurrences in the world. The linking itself was intuitive and quite straightforward to annotators due to the human readable nature of the CE profile and CE steps. Like Provenance Linking, Mention Annotation is not intended to create an exhaustively labeled corpus. Instead, Mention Annotation results in a handful of exemplars for each CE and its steps.

Figure 4 shows a fleshed-out event mention frame based on Example 1 from the Provenance Linking task (in section 3.3.2 above), which is anchored to a video element of a multi-media document in English.

Event Description	Donated items were collected for flood victims in Stainforth		
Anchor	K0C03N803 0:59-1:05		
Type	Transaction.Donation.Unspecified		
Attribute	Actual		
Argument	**Entity/Event**	**Anchor**	**Type**
Giver	N/A		
Recipient	N/A		
Beneficiary	flood victims	41.324-45.639	PER
ArtifactMoney	donated items	39.654-74.589	COM
Place	Stainforth village	57.759-64.239	GPE
Temporal	**determined from document context**		
Start	Before 2019-11-15		
End	unknown		
Link	**CE Step from CE009**		
Step	Step 1.2: Collect Physical Aid		

Figure 4: Event Mention Frame

There may be multiple event or relation mentions of a certain complex event step that occur in the same document. Whether or not to label a particular mention depends on whether it provides novel information about a CE step. To decide whether the information is novel, annotators follow these principles:

- If a single CE step is depicted in two different actions that match two different event types, separate annotations are created for both of those two event mentions, even though they represent the same CE step.
- If mentions of the same or different event or relation types instantiate the same CE step in different data modalities, each of them is annotated separately, as an event mention from a different data modality is considered to be novel information.
- If a mention adds new information (e.g., different arguments about the CE step), it is annotated, even if it covers an event or relation type or data type that has already been annotated.

Entity mentions are annotated only in their occurrence as event or relation arguments, filling argument slots in the event or relation frame. Additionally, within-document coreference of argument entities is provided, when a coreferent entity fills argument slots in multiple events or relations.

Relations between events are not annotated in the SLC, but some event relations may be derivable from the temporal information provided within the event frame annotation, as well as from the linking to CE steps (e.g., preconditions, temporal relations across CE steps), and some events are annotated as the argument of other events (e.g., the topic of a protest event).

In all, we annotated a total of 2499 event and relation mentions in 346 multi-media documents. Table 5 shows the number event and relation mentions annotated for documents in each language in each domain.

Domain	**English Mentions**	**Spanish Mentions**	**Total**
Business Workings	246	49	295
Civil Unrest	183	16	199
Conflict Threat	207	77	284
Cyber Information	86	41	127
Disaster	214	93	307
Government Workings	177	86	263
Illegal Activity	111	31	142
Legal Proceedings	176	40	216
Medical Intervention	108	13	121
Movement	57	2	59
New Capacity Development	262	60	322
Social Life	139	25	164
Total	1966	533	2499

Table 5: Tally of mention annotations for each language in each domain.

3.4 Quality control

Quality control for the SLC corpus included both manual and automatic checks for consistency, completeness and data integrity. In addition to review by SLC project team leaders, an independent quality review was conducted by an external organization. For each CE, external reviewers checked the coverage provided by annotated instances to confirm:

- Sufficient coverage of CE steps, languages, media types, other features is provided by annotated event and relation mentions.
- Every document contains good exemplars of CE/steps for system development.
- As a whole, annotations reflect a good range of diversity of real-world incidents for CE instantiation.
- Event, relation, entity arguments are present in diverse media elements.

Annotation quality was also checked to confirm that for a given category (e.g., type and argument role, among others) all labeled mentions are reasonable. In addition, a quantitative and qualitative review of trends across annotators was conducted, to identify over-/under-tagging tendencies and the use of different tags for the same event or step. For example, this check would flag annotator A who frequently uses a Contact event or annotator B who uses TeachingTrainingLearning for the same concepts.

The issues identified during quality control were folded into the on-going annotator training process to ensure improvement of overall annotation quality for subsequent data.

4. Conclusion

The Schema Learning Corpus represents a new linguistic resource for real-world complex events in multilingual, multimedia data, supporting research into the use of event schemas within systems that can reason about real-world events in order to help users understand how individual events fit together and how events may develop over time. The SLC incorporates large volumes of background data in English, Spanish and Russian, and defines 100 complex events across 12 domains, with CE profiles containing information about the typical steps and substeps and expected event categories for the CE. Multiple documents are labeled for each CE, with pointers to evidence in the document for each CE step, plus labeled events and relations along with their arguments across a large tag set. The SLC comprises over 16.2 million background documents, 3433 documents relevant to complex events, provenance linking annotation for 1171 documents, and event and relation annotation for 346 documents. The Schema Learning Corpus was initially made available to performers within the KAIROS program. It is currently being prepared for publication in the Linguistic Data Consortium catalog, making it broadly available for research.

5. Acknowledgments

This effort was sponsored by the Air Force Research Laboratory (AFRL) and the Defense Advanced Research Projects Agency (DARPA).

The authors also gratefully acknowledge the contributions of annotation coordinators Kira Griffitt, Neil Kuster and Ann O'Brien, technical infrastructure developers Christopher Caruso, Brian Gainor, Jonathan Wright and David Graff, and the work of English and Spanish annotators who contributed to this corpus.

6. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley Framenet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Bejan, C. and Harabagiu, S. (2010). Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422, Uppsala, Sweden, July. Association for Computational Linguistics.

Cybulska, A. and Vossen, P. (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the 9th Language Resources and Evaluation Conference* (LREC2014), Reykjavik, Iceland.

DARPA. (2018). Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS). Defense Advanced Research Projects Agency, DARPA BAA HR001119S0014.

Hovy, E., Mitamura, T., Verdejo, F., Araki, J., and Philpot, A. (2013). Events are not simple: Identity, non-identity, and quasi-identity. In Proceedings of the 2nd *Workshop on events: Definition, detection, coreference, and representation* (pp. 21-28).

Li, M., Xu, R., Wang, S., Zhou, L., Lin, X., Zhu, C., and Chang, S. F. (2022). CLIP-Event: Connecting Text and Images with Event Structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pages 16420-16429, New Orlean, United States.

Mitamura, T., Liu, Z., and Hovy, E. (2017). Events Detection, Coreference and Sequencing: What's next? Overview of the TAC KBP 2017 Event Track. In *Proceedings of TAC KBP 2017 Workshop*.

Piaget, J. (1965). *The language and thought of the child*. New York, NY: Harcourt, Brace & World.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., and Ma, X. (2015). From Light to Rich ERE: Annotation of Entities, Relations, and Events. In Proceedings of *The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Tracey, J., Bies, A., Getman, J., Griffitt, K., and Strassel, S. (2022). A Study in Contradiction: Data and Annotation for AIDA Focusing on Informational Conflict in Russia-Ukraine Relations. In *Proceedings of the 13th Language Resources and Evaluation Conference* (LREC 2022), Marseille, France.

7. Language Resource References

Chen, S., Strassel, S., and Mott, J. 2020. DEFT Chinese Light and Rich ERE Annotation. Linguistic Data Consortium, LDC Catalog No.: LDC2020T19.

Chen, S., Bies, A., Griffitt, K., Ellis, J., and Strassel, S. 2023. DEFT English Light and Rich ERE Annotation. Linguistic Data Consortium, LDC Catalog No.: LDC2023T04.

Mitchell, A., Strassel, S., Huang, S., and Zakhary, R. 2005. ACE 2004 Multilingual Training Corpus. Linguistic Data Consortium, LDC Catalog No.: LDC2005T09.

Walker, C., Strassel, S., Medero, J., and Maeda, K. 2006. ACE 2005 Multilingual Training Corpus. Linguistic Data Consortium, LDC Catalog No.: LDC2006T06.