Spanless Event Annotation for Corpus-Wide Complex Event Understanding

Ann Bies, Jennifer Tracey, Ann O'Brien, Song Chen, Stephanie Strassel
Linguistic Data Consortium
3600 Market Street, Suite 810, Philadelphia, Pennsylvania, 19104, United States
{bies, annsz, zhiyi, strassel}@ldc.upenn.edu

Abstract

We present a new approach to event annotation designed to promote whole-corpus understanding of complex events in multilingual, multimedia data as part of the DARPA Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS) Program. KAIROS aims to build technology capable of reasoning about complex real-world events like a specific terrorist attack in order to provide actionable insights to end users. KAIROS systems extract events from a corpus, aggregate information into a coherent semantic representation, and instantiate observed events or predict unseen but expected events using a relevant event schema selected from a generalized schema library. To support development and testing for KAIROS Phase 2B we created a complex event annotation corpus that, instead of individual event mentions anchored in document spans with pre-defined event type labels, comprises a series of temporally ordered event frames populated with information aggregated from the whole corpus and labeled with an unconstrained tag set based on Wikidata Qnodes. The corpus makes a unique contribution to the resource landscape for information extraction, addressing gaps in the availability of multilingual, multimedia corpora for schema-based event representation. The corpus will be made available through publication in the Linguistic Data Consortium (LDC) catalog.

Keywords: cross-document events, information extraction, multimedia corpora

1. Introduction

In today's complex information landscape, there is a compelling need for human language technology that can help people quickly understand connections between seemingly unrelated events and make predictions about how evolving real-world incidents are likely to unfold in the future. This need is particularly acute when considering the limitations of current technology for multilingual, multimedia data. The DARPA Knowledge-directed Artificial Intelligence Reasoning Over Schemas (KAIROS) Program aims to build technology capable of understanding and reasoning about complex real-world events like a specific terrorist attack or disease outbreak in order to provide actionable insights to end users (DARPA, 2018).

KAIROS systems include 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 complex events within a large multilingual, multimedia corpus. While the number of manually labeled complex events for development and testing is necessarily limited due to resource constraints, KAIROS technologies are required to be able to handle any complex event in the data, regardless of its annotation status. Systems are required to extract events (i.e., real world events that occur in the data set) and their arguments (i.e., the participants in each real world event) from individual multimedia documents, aggregate that information across documents and languages into a coherent semantic representation, and instantiate observed events or predict unseen but expected events for the incident using an abstract event schema selected from a generalized schema library. This task is made all the more challenging when information relevant to obtaining a comprehensive understanding of what happened and what is about to happen is scattered across a document set comprising several different languages, data types and genres.

This paper describes a new approach to event annotation first developed to support KAIROS system development and evaluation. In keeping with the program's goals of achieving whole-corpus event understanding, annotation does not utilize individual event mentions or document spans (i.e., character spans, image bounding boxes, audio timestamps or video frames) to "anchor" event annotations; in fact, no document-level provenance is provided. Instead, annotations consist of a series of event frames, structured representations populated with information aggregated from across the whole data set and temporally ordered relative to one another. This annotation approach of aggregating information from the whole corpus in order to provide a big picture narrative of an incident is better adapted to schema or script-based event representation than a set of fine-grained individual event mentions would be (Chambers and Jurafsky, 2008 and 2009).

The resulting set of labeled events must be comprehensive enough to tell the story of the incident as a whole, but also compact enough to be understood quickly by a human reader. This requires annotators to make challenging decisions about the granularity of each event, as well as deciding how to merge or split closely related events. For example, in a developing disease outbreak incident, it is not necessary to create separate structured events for each individual person who is sick – and in fact, having potentially hundreds of individual sickness events might make overall understanding of the outbreak more difficult. Instead, one coarser-grained illness event that aggregates all of the relevant people, places and times better supports a big picture

LREC-COLING 2024, pages 15105–15113

understanding of the unfolding disease outbreak incident.

During evaluation, KAIROS system output is compared with the set of temporally ordered reference events for each incident through both automatic and manual assessment. The use of coarse-grained events in the reference annotation avoids penalizing systems for failing to identify every individual sickness event in the corpus, but rewards systems that include an aggregated sickness event (with appropriate arguments and temporal information) in their instantiated disease outbreak schema. Evaluation also examines the system's ability to accurately predict events that do not appear in the test set but which have been manually labeled as part of the same complex event incident within a hidden corpus.

The corpus-wide, spanless event annotation approach was applied to 15 multimedia (text, video, image, audio) data sets in English, Spanish and Russian for the KAIROS Phase 2B evaluation cycle. Each unique real-world event present in the corpus was manually annotated to yield a single populated event frame, drawing arguments and event features from across all languages and modalities. Further, instead of applying a single event type from a limited tagset, each event and all event arguments were assigned one or more Wikidata Qnode types. The resulting annotations were used in KAIROS system development and as part of an assessment-based evaluation protocol.

2. Related Work

The KAIROS Phase 2B corpus makes several contributions to linguistic resources, especially due to the multimedia nature of the corpus, and the focus on cross-document coarse-grained event annotation and the linking of both events and event arguments to Wikidata Qnodes in order to understand the relevant events that make up a complex real world incident.

Manual event annotation is frequently done by labeling all mentions of all events that conform to a pre-specified annotation tag set (Doddington et al., 2004; Matsuyoshi et al. 2010 ; Fokkens-Zwirello et al. 2013 ; Walker et al., 2006; Chen et al., 2023). This approach results in spans of text from the document corpus, with each span labeled with an event type tag. However, understanding events across a whole corpus via event mention annotation requires difficult event coreference decisions (Liu et al., 2015; Song et al., 2018) which are made even more challenging when data is multilingual or from multiple modalities including non-text sources.

Multimedia data -- especially documents on the web that contain embedded images, video and audio clips, infographics, social media snippets and reader comments interwoven with traditional text -- is gaining importance as users seek robust technologies that can extract key knowledge elements from diverse information streams. Especially for Russian, but for Spanish and English as well, most event corpora focus on one media type or include multiple media types but treat each media file as a self-contained document. Exceptions are mostly English corpora of subtitled video or images with captions (e.g., Han et al., 2018; Lin et al., 2015), and the multimedia data for English, Ukrainian and Russian in the AIDA corpus (Tracey, et al., 2022). The KAIROS Phase 2B corpus presents a new addition to the available resources by providing a multilingual and multimedia dataset designed to support understanding, event prediction and evaluation of complex events in a holistic way.

Prior mention-based event annotation efforts may also involve manual decisions about event coreference (e.g., Song et al., 2018). The difficulties inherent in determining subevent structure and in resolving the coreference of event mentions are discussed in Araki et al. (2014). The spanless, corpus-wide annotation approach defined by KAIROS removes the need to annotate event coreference, reducing the time needed for annotation and simplifying the annotator decision-making process.

Finally, most prior event mention annotation efforts have utilized a pre-defined, limited annotation tag set that is developed to closely align with the specific domain represented in the corpus being annotated (e.g., Doddington et al., 2004; Mitamura et al., 2015; Chen et al., 2023; Walker et al., 2006). More recent work has begun exploring the use of Wikidata Quodes for event type assignment, via the ontology overlay (Spaulding et al., 2023); this approach provides a much larger, open and domain independent set of event (and entity) types for annotation. GLEN (Li et al., 2023) and UMR (Bonn et al., 2023; Van Gysel et al., 2021) also use PropBank roleset links to Wikidata Qnodes. Our approach to event annotation continues this trend, utilizing Qnodes rather than a fixed annotation tagset.

3. Annotation

Annotated human reference events for each complex event incident (CE) inform automated evaluation and scoring as well as human assessment of system schemas with events extracted or predicted by systems. Annotators create one frame per unique event per CE, synthesizing information from all input documents for the CE. Annotators create the minimal set of events that are needed to describe each incident, with a general preference for fewer coarse-grained events rather than more fine-grained events. Timestamp information is included as part of the event frame for each event; additionally, all events within the CE are temporally ordered in a two-layered relative start order. Inference and logical reasoning are permitted for arguments of events and for temporal information, but events and relations must be explicitly mentioned in the data in order to be labeled in the reference annotation.

Events and relations (including their arguments) that comprise each CE are annotated based on understanding garnered from the input data set as a whole, and as such do not include document provenance such as text offsets or video timestamps. Events and relations that are mentioned in multiple documents or in multiple media types (e.g., text and

video) are represented only once, with their arguments and attributes reflecting information gathered from the full data set. In order to enable manipulation of the data sets into hidden and exposed partitions for certain evaluation conditions, annotators keep track of which documents contain reference to which events, relations and arguments for a given incident.

The event, relation, and entity annotation tagset is based on Wikidata nodes, making use of the Wikidata overlay provided by Spaulding et al. (2023) to allow annotators to select event and relation types and their associated argument roles from a broader set of curated Wikidata nodes.

Wikidata and the overlay provide broad coverage of the events involved in many different kinds of real-world incidents. However, either Wikidata or the overlay may lack good coverage for some scenarios. If no appropriate node can be found in the overlay, specialist annotators search Wikidata directly to find an appropriate node to use as a type, but this occurs only rarely. If searching Wikidata directly does not provide an appropriate scenario-specific fine-grained event type, specialist annotators may select a coarse-grained event type that is appropriate to the scenario and the incident being annotated.

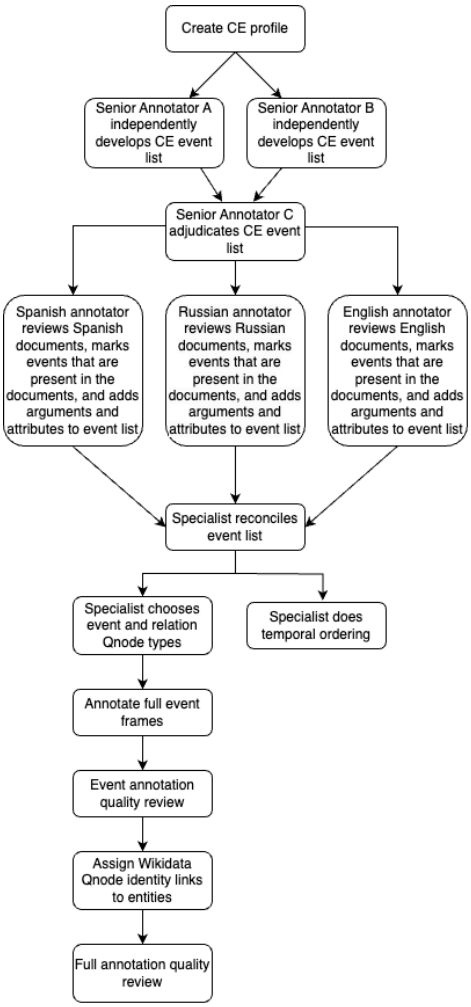

Figure 1. Annotation Workflow for Corpus-wide Spanless Event Annotation

Temporal ordering of all events within each CE is annotated after all event frames for the incident have been created. Temporal ordering indicates the start order for each event, relative to other events for the CE. The event start may be specified as exactly, before or after some numbered order, or may be specified as unknown. Before and after may be combined. More than one event can have the same start order relative to other events. While relations provide contextual information about the CE, they do not in and of themselves constitute steps in the complex event and so are not labeled for temporal ordering.

Annotation of arguments, including linking entities to Wikidata and doing coreference for entities that appear as arguments for multiple events, is completed after all event frames are created. For all entities annotated as the argument of an event or relation, a unique entity ID assigned. All argument entities are linked to Wikidata, a form of Knowledge Base (KB) linking that uses Wikidata as the KB. An identity link is assigned if the entity has a Qnode in Wikidata that uniquely identifies that entity. If no identity link can be found for a given entity, it is assigned a unique NIL ID.

Figure 1 shows the annotation workflow and the steps involved in iteratively developing an event list that contains the minimal set of events needed to fully represent each CE incident, followed by full event frame annotation, temporal ordering annotation, entity linking annotation, and quality review.

One of the most challenging and time-consuming aspects of traditional span-based event annotation is the decision-making process around identifying whether a given event mention is taggable and then determining its type; another is making difficult event coreference decisions (Song et al., 2018). The spanless event annotation approach described here improves the efficiency of event frame annotation by making the difficult decisions about taggability in a holistic way, solely based on relevance to the real-world incident and abstracted away from how the event is expressed in the data. This means that annotators doing event frame annotation can focus on accurate annotation of arguments and properties of the event rather than on worry about whether the event is taggable in the first place. Moreover, because annotators create one frame per unique event per CE, event coreference decisions are not needed. In addition, having a specialist make the Wikidata Qnode type decisions for the refined minimal set of events for each incident prior to event frame annotation, and across all incidents simultaneously, provides an overarching view of the events that allows for consistency in type assignment throughout the corpus. There are potential scaling challenges due to the effort required to develop the minimal necessary set of events for each CE and also make Qnode type decisions across the corpus. However, the resulting cross-document event annotation allows for each event frame to include the richest, most complete information about the event from the document set as a whole, so individual event frames are richer than in

a mention-based approach, reducing the need for effortful exhaustive mention annotation.

3.1 Source Data

Each evaluation in the KAIROS Program is centered around one or more scenarios, defined as an overarching type of complex event, plus real-world incidents that exemplify that kind of event. KAIROS scenarios are complex, encompassing multiple sub-events that can occur sequentially, simultaneously, or unordered, and involving multiple participants. Table 1 includes a complete list of the CEs that were annotated under each scenario in this corpus and also shows the scenario and incident name for each CE in the KAIROS Phase 2B Evaluation.

Scenario	CE_ID	CE Name
Hazardous Spills	CE2201	2014 Chemical spill at Grupo Mexico's Buenavista del Cobre mine in the Mexican state of Sonora
	CE2202	2020 Mauritius Oil Spill
	CE2203	2022 Aqaba gas leak
Riots	CE2204	2013 Little India riot (December 8, 2013 - Singapore)
	CE2205	2021 Dutch curfew riots (23–26 January 2021)
	CE2206	2021 Spain riots about imprisonment of Pablo Hasél (16 February 2021 – 27 February 2021)
Disease Outbreaks	CE2207	2017 Dengue, Peshawar, Pakistan
	CE2208	2017 Marburg, Uganda, October
	CE2209	2018 Nipah Virus, India
Terrorist Attacks	CE2210	Brussels bombing (March 22, 2016)
	CE2211	Nice Truck Attack (July 14, 2016)
	CE2212	Strasbourg Christmas Market Attack (11 December 2018)
Coups	CE2213	2010 Niger coup d'état (18 February 2010)
	CE2214	2021 Guinean coup d'état (5 September 2021)
	CE2215	2021 Myanmar coup d'état (1 February 2021)

Table 1. All Scenarios and Complex Events Annotated with Reference Annotation for KAIROS Phase 2B Corpus

ID: CE2201

Title: 2014 Chemical spill at Grupo México's Buenavista del Cobre mine in the Mexican state of Sonora

Summary of the CE: On August 6, 2014, the Buenavista del Cobre copper mine in Sonora, Mexico, which was owned by Grupo México, spilled 11 million gallons of copper-sulfate acid from its containment pond into nearby waterways. The Bacanuchi and Sonora Rivers were contaminated, and surrounding communities were heavily impacted. The contamination disrupted daily life for hundreds of thousands of residents, with many lacking clean drinking water. Grupo México and environmental authorities cleaned up and remediated the contamination. The spill resulted in a criminal complaint against the Buenavista del Cobre mine, and a fine of 3.4 million pesos was imposed on Grupo México by Mexico's Federal Department of Environmental Protection.

Scope of event: This event begins with the negligence that created the conditions for the disaster. It ends with the imposition of fines as the penalty the company will pay for their role in the spill.

Figure 2. Complex Event Profile: 2014 Chemical spill at Grupo México's Buenavista del Cobre mine in the Mexican state of Sonora (CE2201)

Within each scenario, we select several specific incidents for evaluation and create an input data set for each incident. Each input data set focuses on a single CE and includes relevant documents. We create a CE Profile for each input data set that briefly describes the target CE, including a summary of the CE, the boundaries and scope of the incident for annotation purposes, and some of the expected aspects of the CE that may occur in the input data set. An example of a CE profile is shown in Figure 2.

Each input data set for Phase 2B includes around two dozen documents on average. The data includes English, Russian, and/or Spanish data, with text, image and/or video content, based on data availability. All documents within the input data set are subject to manual annotation. Table 2 shows the document counts and languages included in each CE input data set. Table 3 shows the count of blog posts, text documents (news articles and press releases), video documents and other document types that were annotated as part of the input data set for each CE. Note that many of the documents are multimedia documents that also include images in their content.

CE_ID	English	Russian	Spanish	Total
CE2201	12	0	8	20
CE2202	12	8	11	31
CE2203	9	6	8	23
CE2204	15	3	4	22
CE2205	6	9	11	26
CE2206	7	12	7	26
CE2207	16	0	3	19
CE2208	7	5	7	19
CE2209	13	8	5	26
CE2210	4	7	11	22
CE2211	5	10	4	19
CE2212	7	6	9	22
CE2213	9	2	9	20
CE2214	10	4	9	23
CE2215	7	3	6	16

Table 2. Document Counts by Language in Input Data Sets for Each Complex Event

CE_ID	Other	Blog	Text	Video	Total
CE2201	5	2	7	6	20
CE2202	21	0	5	5	31
CE2203	6	0	14	3	23
CE2204	4	0	11	7	22
CE2205	15	0	7	4	26
CE2206	12	0	8	6	26
CE2207	6	0	10	3	19
CE2208	10	0	6	3	19
CE2209	9	1	14	2	26
CE2210	15	0	6	1	22
CE2211	11	0	6	2	19
CE2212	10	0	11	1	22
CE2213	3	0	14	3	20
CE2214	8	0	13	2	23
CE2215	6	1	4	5	16

Table 3. Document Distribution by Data Type for Each Complex Event

In constructing the input data sets, preference was given to individual documents that covered just a few of the events comprising the incident, as well as to documents that contained unique information relative to the rest of the corpus. Documents reflecting the evolving nature of the incident -- for instance, where the number of victims or the identity of the attacker changes from one document to the next -- were also preferred. Individual documents that gave a retrospective summary of the entire incident and all its

component events were strongly dispreferred. This strategy allowed for better evaluation of KAIROS system capabilities, reflecting how real world incidents unfold over time, and supporting partitioning of the data into seen and unseen documents for prediction evaluation.

3.2 Annotation Principles

Corpus-wide spanless annotation captures relevant information from the whole input data set to characterize the events and relations that are needed to understand the CE. This means that all events and relations, as well as their arguments, attributes, and temporal information are drawn from the entire set of relevant English, Russian, and Spanish documents, rather than from just one document. If a document provides unique information about an event or relation (either attributes or arguments) that is not present in the other documents, that information is also added to the annotation for that event/relation. In this respect, the event and relation annotation is a full representation of the information contained in the on-topic documents.

Only events and relations that are relevant to the CE are annotated. In addition, events that are not themselves relevant to the CE but are required arguments of relevant events/relations are labeled so that they can fill the argument slot needed to complete the annotation frame of the relevant event/relation.

Although annotators do not label individual event and relation mentions, and do not label spans, strings, or offsets, an event or relation must be directly mentioned in the data in order to be taggable. The event or relation can be mentioned in any modality (text, audio, video or image), in any language, in one or more documents in the data set for the CE. Events and relations cannot be inferred; they must be directly observed in the data. The participation of arguments in events however can be inferred.

Example 1:

- Document 1: Shirley Mae Almer died on Sunday after eating contaminated peanut butter in a nursing home in Brainerd.
- Document 2: Relatives are suing a Brainerd nursing home after a 72-year-old woman died on December 21, 2008 of salmonellosis.

This example shows a death event with argument and temporal information spread across multiple documents. The person who died is referred to with a nominal description of “a 72-year-old woman” in one document and also by name in the other document. Only the first document includes the name of the deceased. Only the second document includes the date of the death and the cause of death. When deciding how to annotate the death event annotators incorporate information from both documents to provide the single most specific representation of the event. So for this example, a single death event is annotated, with the deceased person named as *Shirley Mae Almer*, the cause of death argument as *salmonellosis*, and the date of death as *2008-12-21*.

Some subjectivity in event framing is inevitable, as different annotators or end users may have different perspectives on what elements matter most for a given complex event. The KAIROS program dealt with subjectivity through adopting an inclusive "reasonableness" standard for annotation and evaluation: if it’s reasonable to consider an event as crucial to the narrative for the incident, the event should be included.

3.3 Annotation Procedure

3.3.1 Annotation of Events, Relations, and Entities

As a first step, we select a rich summary document about the incident for each CE, such as a Wikipedia page about the incident. Because the input data sets focus on the evolution of the incident in real time as much as possible and typically do not include comprehensive information in a single document, the summary document itself is not part of the input data set. Annotators may refer to the CE Profile as they use the summary document to create an initial list of events that are critical to understanding the CE. Types are not assigned to events at this stage; each event is represented by a sentence-length natural language description that contains all of the information about what happened and the arguments involved (e.g., “The Public Health Rapid Response Team investigated the cholera outbreak in the Dominican Republic”). The focus of the annotation is to produce a complete list of the crucial events that make up the CE. Events that are trivially relevant, but not important to understanding the CE are not annotated.

Native speaker annotators then review all documents in the input data set one at a time to refine the list of events, focusing on the documents in their native language. Annotators read every text document, view every video, and examine every image in the document set for their language, using an annotation interface to view text or images and to play video. At this stage, annotators identify any event, participant, or attribute information that needs to be added or updated in the event descriptions in the initial list of events for the CE in order to fully capture the CE incident. This may include adding arguments, refining timestamp information and/or proposing that additional critical events or relations be added to the event list. At this stage, annotators also verify the presence of events from the summary document in the input data set, and exclude any events or relations from annotation that do not occur in the input data set (even if they may have occurred in the summary document).

The corpus includes English, Russian and Spanish data, and the methods and techniques were the same for all languages. The size of the annotation team varied by language and complex event. English, Russian, and Spanish annotators reviewed the

documents in their own language from each CE's data set and identified the events in the adjudicated event list for the CE that were present in those documents, in addition to refining the events in the event list.

All events or relations in the reference annotation must have direct evidence in the input data set, but annotators may infer the presence of arguments for events based on whole-corpus understanding and a reasonable, intuitive interpretation of the data. That is, annotators may add arguments that are not directly attested in the event argument role in the source data if a reasonable reader would conclude that the argument must have participated in the event or relation. For example, a given input data set may contain multiple events describing how Martin Farnsworth built a homemade bomb and attempted to set it off at Pine View High School. Several documents state that the bomb went off, but they don't explicitly say that Martin was the one who detonated it. However, annotators can use reasonable inference to add Martin Farnsworth as the agent argument for the detonate event if that is their understanding of what happened after inspecting the input data set.

Annotators annotate the most complete realization of each event or relation, incorporating all of the information from their understanding of the whole input data set. Annotation reflects the final and most specific state of affairs based on the information in the data set. For instance, if an early document reports that 418 persons were ill and later documents indicate that there were 647 illnesses, the annotation will include one illness event with an argument of 647 (not 418) victims.

When annotating an event or relation, annotators fill in the frame with as much information as possible about the event from the corpus, pulling information about the event or relation and its arguments from the whole document set, regardless of language or modality. Arguments can be drawn from anywhere in the document set - from any document, data modality, or language. All arguments that add information to the event are included, and multiple arguments can be added to the same argument role. Arguments may themselves be events, relations, or entities. The most specific, accurate, and recent piece of information about the argument that appears in the document set is used to label each argument. Events and relations are not associated with any specific document provenance, and no span or offset justification is provided. Events and relations that are mentioned in multiple documents are listed only once, with their arguments and attributes reflecting the information available in the input data set as a whole.

Once the list of events is final, including all arguments, expert annotators assign a Qnode type to each event in the list, making use of the DWD event overlay produced by the Cross-Program Ontology Working Group (Spaulding et al., 2023). All arguments are assigned role labels based on the Qnode type selected, using the role labels in the overlay. Annotators also assign Qnode types and argument roles to relations.

An example of a fully annotated event frame is shown in Figure 3. The completed event frame consists of a natural language description, Qnode type and timestamp, along with the participating arguments, their Qnode types, and their roles in the event. Each event in the CE is fully annotated in this way.

Event			
Description	Peanut butter from Peanut Corporation of America was contaminated with salmonella		
Type	Q60528603 contamination		
Attribute	none		
Timestamp	Start before 2008-10-21TXX:XX:XX End before 2008-10-21TXX:XX:XX		
Argument	**Type Qnode**	**Role**	**Attribute**
Peanut Corporation of America	Q43229 organization	A0-pag-agent-causer_of_contamination	none
Peanut butter	Q15401930 work product	A1-gol-destination-thing_becoming_contaminated	none
Salmonella	Q2826767 disease causative agent	A2-ppt-theme-contaminant	none

Figure 3. Annotated Event Frame

3.3.2 Linking Events and Entities to Wikidata

Event types are assigned as Wikidata Qnodes, with associated role labels from the overlay as above. In addition, entities that appear as arguments in the final list of events for each CE are also linked to Wikidata Qnodes. During this stage of annotation, annotators assign fine-grained entity types to all arguments, using the most specific Wikidata Qnode they can find for each entity. In addition, entities that can be linked to a specific, unique identity node (e.g., "CDC" would be linked to "Centers for Disease Control and Prevention" Q583725) receive an identity ID for that Qnode; unique NIL IDs are assigned to entities for which an appropriate identity link cannot be found in DWD. Entities may appear as arguments in multiple events or relations, and when the same entity appears with multiple types across arguments they are labeled as coreferent. (See Figure 3 for an example of assigned Wikidata Qnodes for an event and its arguments.)

Because the annotation is based on information drawn from multiple documents, multiple languages, and multiple modalities, annotators provide a brief natural language description of each entity written in English. This description combined with the identity Qnode and the fine-grained type Qnode completes the entity annotation. Because we are not annotating each mention in the corpus separately, there are no spans or offsets associated with the entity. Entities are only annotated when they serve as an argument to one or more events or relations that are annotated for the CE.

3.3.3 Temporal Start Order

In addition to labeling specific timestamp information for each event frame, the final list of annotated events for each CE is also put in temporal order based on start order, using the annotator's understanding and logical reasoning about the order in which events started during the incident. For aggregated coarse-grained events, annotators order by the earliest start time they understand in the aggregate. Annotating

only events that are critically relevant to the CE and annotating coarse-grained events allows the temporal start order to be relatively straightforward.

Once the full set of event frames has been created for the input data set, annotators indicate temporal ordering by specifying the start order for each event, relative to other events for the CE. The event start order may be specified as exactly, before or after some numbered order, or may be specified as unknown. Before and after may be combined. More than one event can have the same start order relative to other events.

For instance:

- Event A order "exactly 6" means that event A started after step 5 and before step 7
- Event B order "exactly 6" and Event C order "exactly 6" means both B and C started after step 5 and before step 7
- Event D order "before 6" means D started sometime before step 6
- Event E order "after 6" means E started sometime after step 6
- Event F order "after 6, before 21" means F started sometime after step 6 but before step 21 (i.e., it started between steps 7 and 20, inclusive)
- Event G order "unknown" means the start ordering of G with respect to the start of the other events is unknown

While relations provide contextual information about the CE, they do not in and of themselves constitute steps in the complex event and so are not labeled for temporal ordering.

Table 4 shows temporal start order annotation for a subset of the events for one CE.

3.3.4 Measuring IAA for Spanless Annotation

The annotation described here differs from previous event annotation in a number of ways that affect Inter-annotator agreement (IAA) scoring (non-span, non-exhaustive, Qnode labels), and as a result, previous methods to measure IAA cannot be directly applied in this case.

We developed a new method to compare spanless annotations for this corpus that relies on defining a set of attributes for each annotated element and a weighted comparison of each attribute. We also manually scored 1021 event pairs to evaluate the automatic method. Using this method, we achieved an f-score of .75 when comparing event annotations across annotators.

Event List for CE2201 - Grupo Mexico Spill	**Temporal Start Order**	
Event Descriptions	**First Layer Start Order**	**Second Layer Start Order**
Grupo México did not study environmental impacts in Sonora before August 6 2014	Exactly 1	n/a
Buenavista del Cobre mine is a subsidiary of Grupo México	n/a (relation)	n/a
There was heavy rain in Sonora on or before August 6 2014	Exactly 2	n/a
11 million gallons of toxic copper-sulfate acid poured out of a containment pond at the Buenavista del Cobre mine into the Tinajas stream and the Bacanuchi and Sonora rivers on August 6 2014	Exactly 3	n/a
11 million gallons of toxic copper-sulfate acid contaminated the Bacanuchi and Sonora rivers, the Tinajas stream, and the El Molinito reservoir on August 6 2014	Exactly 4	n/a
Grupo México did not report the spill to authorities on August 6 2014 in Sonora	Exactly 5	n/a
More than 800,000 residents, Sindicato Nacional de Mineros Seccion 65 (Section 65 National Miners Union), Ignacio Sanchez Santa Rosa, the Garcia family, ranchers, and almost 7,000 farmers were affected by the spill in Sonora including specifically Hermosillo, Arizpe, Banamichi San Felipe de Jesus, Aconchi, Baviacora and Ures on August 6 2014	Exactly 5	n/a
Some residents of Sonora, including Luz Apocada, Oscar, and 19 other people, had skin rashes after August 6 2014	After 4, Before 7	n/a
11 million gallons of toxic copper-sulfate acid dyed the water in Bacanuchi and Sonora rivers orange after August 6 2014	After 4, Before 7	n/a
Many crops, fish, and livestock were poisoned by the contaminated water in Sonora after August 6 2014	After 4, Before 7	n/a
The water in Bacanuchi had a bad odor after August 6 2014	After 4, Before 7	n/a
More than 300 wells were shut down in Sonora after August 6 2014	After 4, Before 7	Exactly 1
Most wells in Sonora were reopened after August 6 2014	After 4, Before 7	Exactly 2
88 schools in Sonora were closed from August 7 2014 until at least August 14 2017	Exactly 6	n/a

Table 4. Annotated Event Descriptions and Temporal Start Ordering, 2014 Chemical spill at Grupo México's Buenavista del Cobre mine in the Mexican state of Sonora (CE2201)

4. Results and Annotated Corpus

The final annotated corpus consists of 15 annotated document sets, covering three real-world incident CEs for each of five scenarios. The corpus covers incidents in English, Russian, and Spanish, and includes text, image, and video data with an emphasis on data that mimics real-time situation monitoring, including real-time unfolding events, partial information, and social media. The structured annotation for each CE covers all critically relevant events, relations, and arguments, with event and entity linking to Wikidata Qnodes and temporal ordering of all events.

This corpus supported the KAIROS Phase 2B evaluation of performer systems by serving as input to the automated scoring and by serving as the human reference annotation that systems were compared against in manual assessment. Because the annotated events in this corpus are coarse-grained as well as corpus-wide and non-span-based, human assessors were able to make judgments about matching a wide range of system events to the human annotated events. Table 5 shows the total number of critical events annotated for each CE's input data set, along with the total number of unique arguments annotated in each CE.

CE_ID	Annotated Event Count	Annotated Argument Count
CE2201	23	68
CE2202	20	58
CE2203	23	58
CE2204	24	48
CE2205	17	73
CE2206	17	87
CE2207	16	55
CE2208	23	50
CE2209	13	52
CE2210	15	43
CE2211	24	55
CE2212	28	55
CE2213	21	38
CE2214	17	35
CE2215	15	49

Table 5. Number of Reference Events, Relations, and Arguments Annotated for Each CE

The annotated corpus was used in evaluation of KAIROS systems in two ways: first, automatic scoring of systems was based on comparison with the annotated reference events, and second, human assessors compared system output with the annotated reference events during manual assessment. Systems were evaluated on identifying events from the data in the system's instantiated schema for the CE, on their ability to place the events in a reasonable hierarchy, and also on their ability to accurately predict events that do not appear in the test data. Documents containing some of the manually labeled reference events were hidden from the test corpus, which allowed a direct assessment of the system's ability to predict events that had been manually annotated.

The scoring and assessment of system produced and instantiated events based on reference annotation during the previous evaluation in KAIROS Phase 2A supported the further development of systems from Phase 2A into Phase 2B. KAIROS evaluation results are not yet publicly available.

In Phase 2B, the reference annotation, along with annotation-based evaluation and manual assessment of system produced and instantiated hierarchies and predicted events, introduced novel annotation and assessment methods, and these is expected to support ongoing research into the future.

5. Conclusion

Our approach to annotating corpus-wide non-span-based coarse-grained events resulted in richer and more efficient annotation of the 15 data sets in this corpus than the traditional span-based approach to fine-grained event/relation annotation used in much of the prior work discussed in Section 2. The event annotations for each CE serve as a cross-document summary of the incident CE in the form of structured event representations. This makes them particularly suitable for assessment of system events instantiating system schemas.

The KAIROS Phase 2B annotation corpus makes a unique contribution to the available resources for multilingual, multimedia information extraction, with a particular emphasis on cross-document event and schema understanding for specific real world incidents. The corpus will be made publicly available through publication in the Linguistic Data Consortium (LDC) catalog after the conclusion of the KAIROS research program.

6. Acknowledgements

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 and Joshua Parry, technical infrastructure developers Chris Caruso, Brian Gainor, Jamie Strausbaugh, Jonathan Wright and David Graff, Seth Kulick for IAA methods development, and the work of all of the English, Spanish, and Russian annotators who contributed to this corpus.

7. Bibliographical References

Araki, J., Liu, Z., Hovy, E., and Mitamura, T. (2014). Detecting Subevent Structure for Event Coreference Resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bonn, J., Myers, S., Van Gysel, J., Denk, L., Vigus, M., Zhao, J., Cowell, A., Croft, W., Hajič, J., Martin, J., Palmer, A., Palmer, M., Pustejovsky, J., Urešová, Z., Vallejos, R., and Xue, N. (2023). Mapping AMR to UMR: Resources for Adapting Existing Corpora for Cross-Lingual Compatibility. In *Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT,*

GURT/SyntaxFest 2023), pages 74–95, Washington, D.C.. Association for Computational Linguistics.

Chambers, N. and Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Chambers, N. and Jurafsky, D. (2009). Unsupervised Learning of Narrative Schemas and their Participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

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

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Fokkens-Zwirello, A. S., van Erp, M. G. J., Vossen, P. T. J. M., Tonelli, S., Van Hage, W. R., Serafini, L., Sprugnoli, R., and Hoeksema, J. E. (2013). "GAF: A grounded annotation framework for events." In *Proceedings of the first Workshop on EVENTS: Definition, Detection, Coreference and Representation*, pp. 11-20.

Li, S., Zhan, Q., Conger, K., Palmer, M., Ji, H., and Han, J. (2023). GLEN: General-Purpose Event Detection for Thousands of Types. https://arxiv.org/abs/2303.09093, accessed 10/13/2023.

Han, Q., John, M., Kurzhals, K., Messner, J., and Ertl, T. (2018). Visual interactive labeling of large multimedia news corpora. In *Proceedings of Leipzig Symposium on Visualization in Applications (LEVIA'18)*, Leipzig, Germany, October.

Lin, T., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common objects in 1838 context. https://arxiv.org/abs/1405.0312v3, accessed 1/13/2022.

Liu, Z., Mitamura, T., and Hovy, E. (2015). Evaluation Algorithms for Event Nugget Detection : A Pilot Study. In *Proceedings of The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 53–57, Denver, Colorado. Association for Computational Linguistics.

Matsuyoshi, S., Eguchi, M., Sao, C., Murakami, K., Inui, K., and Matsumoto, Y. (2010). "Annotating event mentions in text with modality, focus, and source information." (2010). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*.

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., and Strassel, S. (2015). Event Nugget Annotation: Processes and Issues. In *Proceedings of The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 66–76, Denver, Colorado. Association for Computational Linguistics.

Song, Z., Bies, A., Mott, J., Li, X., Strassel, S., and Caruso, C. (2018). Cross-Document, Cross-Language Event Coreference Annotation Using Event Hoppers. *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Spaulding, E., Conger, K., Gershman, A., Uceda-Sosa, R., Brown, S., Pustejovsky, J., Anick, P., and Palmer, M. (2023). The DARPA Wikidata Overlay: Wikidata as an ontology for NLP, *ISA 2023 at IWCS 2023*.

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 Thirteenth Language Resources and Evaluation Conference*, pages 1831–1838, Marseille, France. European Language Resources Association.

Van Gysel, J., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C., Hajič, J., Martin, J., Oepen, S., Palmer, M., Pustejovsky, J., Vallejos, R., and Xue, N. (2021). Designing a Uniform Meaning Representation for Natural Language Processing, *Künstliche Intelligenz*.

Wikidata. https://www.wikidata.org.

8. Language Resource References

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.

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