EDEN: A Dataset for Event Detection in Norwegian News

Samia Touileb[1], Jeanett Murstad[2], Petter Mæhlum[3], Lubos Steskal[4], Lilja Charlotte Storset[3], Huiling You[3], Lilja Øvrelid[3]
[1]University of Bergen, [2,3]University of Oslo, [4]TV 2
[1]samia.touileb@uib.no, [2]jeanmurs@gmail.com, [3]{pettemae, liljacs, huiliny, liljao}@ifi.uio.no,
[4]lubos.steskal@tv2.no,

Abstract

We present EDEN, the first Norwegian dataset annotated with event information at the sentence level, adapting the widely used ACE event schema to Norwegian. The paper describes the manual annotation of Norwegian text as well as transcribed speech in the news domain, together with inter-annotator agreement and discussions of relevant dataset statistics. We also present preliminary modeling results using a graph-based event parser. The resulting dataset will be made freely available for download and use.

Keywords: event extraction, corpus, annotation, Norwegian, news domain

1. Introduction

Event extraction is a central task in Natural Language Processing (NLP) which enables applications aiming to extract and aggregate information about real world events from texts. Annotations of these events typically provide information about the specific type of event (*e.g.* an attack, injury, or transfer of money) and the participants involved in certain roles (*e.g.*, the agent or victim of an attack, the recipient of a transaction, *etc.*). While several event annotated datasets have been created for a number of languages, no such openly available dataset currently exists for Norwegian.

This paper presents EDEN[1]: the first event-annotated dataset for Norwegian, focusing on the news domain and including both edited news articles as well as transcribed spoken news broadcasts. The annotations adapt the widely used ACE (Automatic Content Extraction) guidelines (Doddington, 2005) to the Norwegian news domain. EDEN (which stands for Event DEtection for Norwegian) contains a total of 630 documents, annotated for 5,805 events and 9,299 arguments.

In the following we briefly review related work before we present the annotation effort, detailing the data sources, pre-annotation procedure, the annotation guidelines with specific adaptations made for Norwegian news text and transcribed speech, as well as inter-annotator agreement scores. We further summarize dataset statistics and present modeling results using a state-of-the-art event parser.

[1]https://github.com/ltgoslo/Event-Detection-for-Norwegian-EDEN-

2. Related Work

2.1. Datasets

One of the first annotation efforts in this area was the Automatic Content Extraction (ACE) program (Doddington et al., 2004) which resulted in richly annotated datasets including entities, relations, and events for English, Arabic, and Chinese. The English ACE dataset comprises 8 event types and 33 specific subtypes. Events in ACE are defined as distinct happenings that entail the involvement of particular participants (Doddington, 2005). An occurrence of such a happening is expressed using an event trigger. Event triggers in ACE are mainly single words, and mostly main verbs (Aguilar et al., 2014). The fundamental objective of tagging an event is therefore to identify and describe events that incorporate various elements participating in different roles in the event, adding detailed information about the occurrence of the event. These participating elements are referred to as event arguments (Doddington, 2005).

The ERE (Song et al., 2015) dataset, also referred to as Light ERE, is very similar to ACE, and further extends on the same annotation scheme to English, Chinese, and Spanish. It also comprises the same event types and subtypes as ACE. However, one of the primary distinctions between these two datasets is the degree of specificity. While annotations in ACE are fine-grained, ERE adopts a more simplified scheme by consolidating tags (Aguilar et al., 2014). ERE also comes in a richer format, dubbed Rich ERE (Song et al., 2015), which extends the Light ERE and comprises 9 event types and 38 event arguments (You et al., 2023).

LREC-COLING 2024, pages 5495–5506
20-25 May, 2024.

2.2. Event Extraction Models

Event extraction is conventionally approached through supervised classification, although alternative methodologies such as generation-based techniques (Paolini et al., 2021; Lu et al., 2021; Li et al., 2021; Hsu et al., 2022), or those influenced by natural language understanding tasks through prompt tuning (Shin et al., 2020; Gao et al., 2021; Li and Liang, 2021; Liu et al., 2022) are gaining prominence. Classification-based approaches decompose event extraction into specific subtasks: trigger detection and classification, as well as argument detection and classification. These subtasks are either addressed individually in a sequential, pipeline-based fashion (Ji and Grishman, 2008; Li et al., 2013; Liu et al., 2020; Du and Cardie, 2020; Li et al., 2020) or jointly inferred as multiple subtasks (Wadden et al., 2019; Lin et al., 2020). For joint modeling, graph-based approaches have recently been proposed, formulating the event extraction task as a structured prediction task. The work of You et al. (2022) presents an adaptation of the PERIN semantic parser (Samuel and Straka, 2020) to the event extraction task. Their work showed promising results on the ACE dataset and is further extended by You et al. (2023) to perform the task of joint information extraction, covering both entities, events and relations derived from different datasets, and handling three distinct languages.

3. Annotation

We here present details on the annotation of the EDEN dataset, summarizing the data sources and pre-annotation, the annotation guidelines, the annotation procedure as well as inter-annotator agreement scores for the annotation effort.

3.1. Data Source

EDEN supports event extraction from news data both for written news and broadcast news. We describe both sources in what follows.

News Text We use the news portion of the Norwegian Dependency Treebank (NDT) (Solberg et al., 2014; Øvrelid and Hohle, 2016), which in addition to having morphosyntactic annotation, has previously been further extended with annotations for named entities (Jørgensen et al., 2020) as well as co-reference information (Mæhlum et al., 2022).

As original category metadata was not present for the newspaper articles in NDT, they were manually annotated with categories from the IPTC Media Topic NewsCodes, which are available for Norwegian[2]. In cases where placement was difficult, more than one category was assigned.

TV News Transcripts In addition to working with standard newspaper articles, we also use a dataset consisting of automated transcripts of the Norwegian television news channel *TV 2 Nyhetskanalen* provided to us by the broadcaster. This dataset contains a sample of transcripts produced between the years 2021 and 2023. The transcripts were generated using a combination of a publicly available commercial third party Automatic Speech Recognition (ASR) provider[3] and an internal processing pipeline. As such, some ASR errors were introduced in the process and their nature can vary over time, as the underlying transcription technology underwent rapid development in the time period.

Due to challenges arising from the use of ASR, and the continuous generation of transcripts over the selected time period, the provided transcripts not only contained news broadcasts but also included all advertisements aired during that time frame. Therefore, we decided to initiate a manual inspection and rectification of these transcripts before selecting documents for annotation purposes.

The transcripts spanned hourly intervals for each day between 2021 and 2023. Our selection criteria led us to focus on a single broadcast per day, specifically those scheduled for 8:00 PM each day. This choice was guided by our informed judgment, grounded in the expectation that the news content during this particular time slot would comprehensively cover the salient events of the day.

In the subsequent phases of our data preparation, we followed several key steps. Firstly, we randomly selected a limited subset of the transcripts, representing each year within the time frame of 2021 to 2023. This resulted in 68 transcripts, which were subsequently manually curated to remove all instances of advertising content.

Furthermore, in order to facilitate the annotation process, we manually divided each transcript into files corresponding to distinct types of broadcasts featured in the 8:00 PM slot. This division was made based on observations during our manual inspection of the transcripts. There was a clear distinction between the different categories of broadcasts on the news channel. For this, we did a simple paragraph separation delimited by two consecutive empty lines, enabling an automated partitioning of each transcript into sub-documents reflecting the diversity of shows and broadcasts. This yielded a selection of 587 distinct documents.

[2] https://iptc.org/std/NewsCodes/mediatopic/treeview/mediatopic-no.html

[3] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-to-text

LIFE	CONTACT	PERSONELL	MOVEMENT	CONFLICT
BE-BORN	MEET	START-POSITION	TRANSPORT	ATTACK
MARRY	PHONE-WRITE	NOMINATE		DEMONSTRATE
DIVORCE		ELECT		
INJURE		END-POSITION		
DIE				
BUSINESS	**TRANSACTION**		**JUSTICE**	
START-ORG	TRANSFER-MONEY	CHARGE-INDICT	FINE	SUE
MERGE-ORG	TRANSFER-OWNERSHIP	ARREST-JAIL	TRIAL-HEARING	CONVICT
	DECLARE-BANKRUPTCY	SENTENCE	RELEASE-PAROLE	AQUIT
	END-ORG	PARDON	APPEAL	EXTRADITE
		EXECUTE		

Table 1: The 8 event categories and the 33 event types present in the ACE dataset.

To further enhance the dataset's quality, we performed an additional curation step. All instances of verbatim duplicated news items within a single day were systematically removed. This step only targeted instances where identical news content was presented in a verbatim manner. Duplicates in which the news item was described with distinct wording or phrasing that might be caused by updates in the event course, were retained. The resulting TV 2 dataset comprises 294 documents.

This data might contain transcription errors, and therefore be of lower quality than the NDT dataset, but we still believe that it is a valuable resource. Including this data provides a different textual modality that allows for event annotation, and extraction, in transcribed spoken news broadcasts.

3.2. Pre-annotation

Since the precise delimitation of argument entity spans can be a source of disagreement during annotation, we make use of either existing syntactic annotation (for the NDT data) or automatically syntactic parses[4] (for the transcribed speech) to provide pre-annotation of potential argument entities. More precisely, we are interested in locating noun phrases that are potential event arguments.

To this end we formulate a set of heuristics over parts-of-speech and dependency relations from the dependency syntax of the treebank. Using the dependency syntax, we extract all nominal heads that are either i) nouns (common or proper nouns), ii) referential personal pronouns[5], iii) possessive pronouns, or iv) adjectives in a nominal syntactic function (subject, object, or prepositional complement). The noun phrase is constructed by traversing all syntactic dependents of these nominal heads. The annotators are instructed to treat the pre-annotated markables as suggestions only, since these NPs will often not correspond to an event argument.

3.3. Annotation guidelines

The annotation work in this project takes as its starting point the English ACE guidelines (Doddington et al., 2004)[6]. These have been widely used and provide a starting point for other event annotation projects (Song et al., 2015; Pouran Ben Veyseh et al., 2022). We here discuss the most essential aspects of the ACE annotation schema to understand the EDEN annotation process, we however focus mainly on the adaptations that were made to Norwegian. We further provide a description of some of the challenges encountered when annotating the second part of our dataset, consisting of automatically transcribed broadcast news.

3.3.1. Event types

An event consists of an event trigger and its arguments. Event arguments are made up of pre-annotated entities (*i.e.* NPs or single tokens). We do not annotate at a subtoken level. In EDEN we operate with a total of 33 different event types, categorized under 8 broader event categories (LIFE, CONTACT, PERSONELL, MOVEMENT, CONFLICT, BUSINESS, TRANSACTION, and JUSTICE) as predefined in ACE. An overview of all the event types can be seen in Table 1.

3.3.2. Event triggers

The event trigger is the text span that most clearly describes an occurrence. We permit triggers to generate multiple events. In most cases, the event trigger will be the main verb of a sentence, as seen

[4]We use the Norwegian UD models provided in Stanza (Qi et al., 2020) to automatically generate the syntactic annotation of the transcriptions.

[5]The NDT annotation identifies so-called formal subjects/objects, which are non-referential or expletive uses of the pronoun *det* 'it'.

[6]https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

in example (1). Depending on the expression of the event, the trigger can also be a noun, participle, or adjective, as shown in example (2) and example (3), respectively. Example (3) shows the participle *skadde*, 'injured' in a predicative position, but participle and adjective triggers can also be found in attributive positions. Example (3) further illustrates an event where a single sentence contains multiple event types: *Krigen*, 'the war', triggers an ATTACK-event, while *skadde* triggers an INJURE-event.

(1) *Hun* ***ringte*** *sønnen sin*
she called the.son her
'She called her son'

(2) *Byen ble rammet av et israelsk* ***angrep***
the.city was affected by an Israeli attack
'The city was affected by an Israeli attack'

(3) ***Krigen*** *etterlot mange* ***skadde***
the.war left many injured
'The war left many injured'

Annotators were instructed to choose minimal spans for potential event triggers, if possible limiting the trigger to a single token. Even so, in certain cases we allow multi-token triggers for an event where no shorter alternatives are available, for instance for particle verbs, as described in Section 3.3.5 below.

3.3.3. Event arguments

Event arguments denote participants involved in an event, described by an event, and attributes (TIME-ARG, PLACE-ARG) that are present in the same sentence as the event trigger (*i.e.* the event's scope). Participant arguments are filled by entities of proper names, pronouns, and other terms referring to a person, organization, or GPE (Geo-Political Entity) depending on the event type. Question words like *hvem*, 'who' are not accepted as participant arguments for events.

Each event type has a set of argument roles that can be filled by entities present in the event's scope, as specified by the ACE guidelines. For the majority of event types, the argument roles are not required to be filled for the event to be annotated. The PHONE-WRITE event is an exception, where two explicit parties (ENTITY-ARG) have to be present in the sentence for the event to be annotated. While we do require two named participants (*e.g.*, *hun*, 'she' and *sønnen sin*, 'her son' in example 1) to annotate PHONE-WRITE, we do not need both to be annotated as ENTITY-ARGs (*i.e.* the communicating agent). That is, it is acceptable to have an ENTITY-ARG role empty when a participant exists as a subtoken of the event's trigger (*e.g.*, *presse-* being a participant unavailable for annotation in the PHONE-WRITE event *Det skriver politiet i en* ***pressemelding***, 'That is what the police are writing in a **press release**').

A particular entity can not fill more than one argument role for one and the same event. For entities such as GPEs, that in principle can be understood as both a place (*i.e.* PLACE-ARG) and participant (*e.g.*, PERSON-ARG, TARGET-ARG) with respect to an event, disambiguation is necessary. The entity takes on the argument role that appears most prominent and relevant given the current event. Disambiguation is done individually for each annotated occurrence. Thus, the role of a particular entity does not have to remain the same for different events within a scope. Since a given sentence may contain more than one event, an entity will, however, often be an argument in several events. In Figure 1, for instance, the entity *Paul Gascoigne* is both PERSON-ARG for the ARREST-JAIL-event and ATTACKER-ARG to the ATTACK-event.

For PERSONELL-events where the position being filled or vacated (POSITION-ARG) includes a GPE (*e.g.*, *sørkoreanske* 'South Korean' in *den* ***tidligere*** *sørkoreanske presidenten*, 'the **former** president of South Korea') the GPE entity will, as a rule, be annotated as ENTITY-ARG (the employing agent) and not PLACE-ARG.

3.3.4. Event modality

In addition to event triggers and arguments, the ACE guidelines annotate additional attributes expressing the *polarity, genericity, tense* and *modality* of the event. We limit our annotation to the *modality* attribute of an event. An event's modality is considered asserted if it is clearly described as having found place, or as currently on-going. Such events are marked with MODAL-ASSERT. All other events, are considered un-asserted, and remain unmarked. Figure 2 shows an example annotation of an asserted event. The sentence describes an arrest that has clearly occurred, and therefore is marked with MODAL-ASSERT, in the shape of an asterisk after the name of the event type.

While all annotated events are unmarked by default, several different types of un-asserted events were defined in ACE. One of these types are fictional events, *e.g. En dag overfaller de en familie på vandring*, 'One day, they assault a family on a hike'. Without context, this sentence can be interpreted as a real occurrence, but it is extracted from a movie review, describing the fictional plot. Therefore, the sentence is not marked as asserted. For other sentences, the un-asserted modality is clear without taking context into account, as in the sentence *Høyre kan danne regjering med FrP*, 'Høyre may form government with Frp'. Figure (3) provides an example of an un-asserted START-ORG

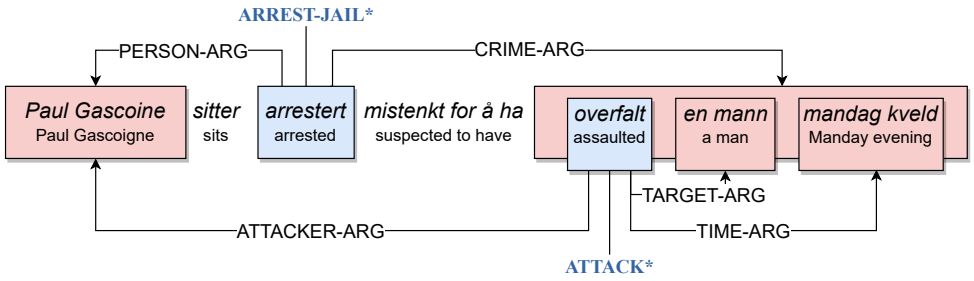

Figure 1: EDEN annotation showing an example of multiple events within a scope, that share an argument entity.

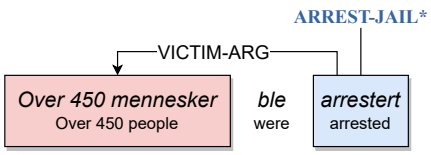

Figure 2: Modality asserted event annotation (indicated by asterisk).

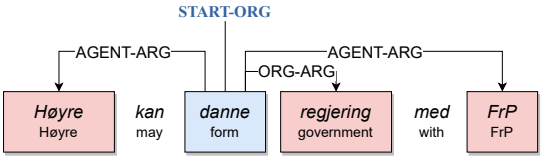

Figure 3: Un-asserted modality annotation.

event, where the phrase *kan danne*, 'may form', clearly expresses the occurrence as hypothetical through the use of a modal verb.

Generic and negated events While we chose to not annotate events with polarity and genericity attributes, it was necessary to decide how generic and negated occurrences of events should be handled. In EDEN, the un-asserted set of events was expanded to include both generic and negated events. Negated events (*e.g.*, the TRANSPORT event *Ingen av dem vil **flytte** fra Norge*, 'None of them wants to **move** from Norway' where the ARTIFACT-ARG *Ingen av dem* negates the event) do not actually take place, and thus fit well into the un-asserted category.

In cases where an event is taking place, but the presence of a potentially participating entity is negated (*e.g.*, *Cecilia Brækhus* in the MEET event: *Cecilia Brækhus kommer ikke til å bokse på det store **VM-stevnet** i Zürich 20. desember*, 'Cecilia Brækhus will not be boxing at the big **world championship event** in Zürich on the 20th of December') we refrain from annotating the entity as a participating argument. Generic events (*e.g.*, *én av 5 sykepleiere **slutter** de første årene av sin yrkeskarriere*, 'one in five nurses **quit** within the first five years of their working career') also differ from asserted events in not being tied to any particular time and place, and can appear similar to the hypothetical event in Figure 3. There are also events that are generic because a potential occurrence is referred to in generic terms, and not as events that actually occur (*e.g.*, the INJURE events in *Årsakene til hukommelsestap kan være mange: **hjernerystelse**, **slag** mot hodet,..*, 'The causes of memory loss can be many: **concussion**, **blow** to the head,..'. This kind of generic events are generally only consisting of a trigger (*e.g.*, **hjernerystelse**), as no potential argument entities exist within the scope.

3.3.5. Adaptations of ACE guidelines

In the following paragraphs we discuss more specific adaptations of the ACE guidelines to Norwegian news texts. Our adaptations consist of both further development of existing ACE guidelines and separate decisions made in this project in order to better capture events in Norwegian news texts.

More concretely, the adaptations were motivated by: i) clarifying parts of the ACE annotation guidelines that were unclear, ii) limiting the task so as to make it manageable and provide consistent annotations, and iii) adapting the guidelines to Norwegian language and society.

Argument selection As a general rule we always choose the entity closest to the trigger when annotating event arguments and attributes. When there are two potential, co-referring argument entities at the same distance from the trigger the longest of the two (*i.e* the one with most tokens) is chosen to fill the argument role since it is assumed to be more informative. If the event trigger is a reflexive verb (*e.g.*, *trekke seg* 'withdraw') the participant-argument will always be a reflexive pronoun, given that it exists in the event's scope (*seg* in the END-POSITION event *Eva Kristin Hansen **trakk** seg som stortingspresident* 'Eva Kristin Hansen **withdrew** from her position as president of parliament').

TIME arguments In order to simplify the annotation task, we operate with only one overarching TIME-ARG attribute instead of the seven time-roles defined in the ACE Timestamp Guidelines (Doddington, 2005). We do not allow for expressions of duration (*e.g.*, *seks dager* 'six days') to be annotated as TIME-ARG for any of the event types. Both multi-word phrases (*e.g.*, *fem år siden* 'five years ago') and simple time expressions (*e.g.*, *mandag* 'Monday' or *2013*) can fill the argument role. With phrases that express time, we try to maintain a minimal argument span and do not include functional elements such as prepositions (*e.g.*, *for* in *for 7 timer siden* '7 hours ago').

When no time expressions (*e.g.*, *uke* 'week', *kveld* 'evening') are present in the scope, other entities that contribute to the positioning of the event on a shared general timeline can be annotated as TIME-ARG. Phrases that relate an occurrence to significant events (*e.g.*, *etter den franske revolusjonen* 'after the french revolution'), or a restricted period of time (*i sommerferien* 'in the summer vacation') have this function. Phrases that relate events only to individual or personal timelines (*e.g.*, *etter jobb* 'after work') do not contain enough information to be annotated as TIME-ARG. For example the phrase *Etter den franske revolusjonen* indicates the event took place in 1799 or later, while *etter jobb* could refer to any hour of any day.

Transfer of ownership The TRANSFER-OWNERSHIP event, according to the ACE guidelines, is only triggered when the ARTIFACT-ARG being transferred is a vehicle, facility, organization, or weapon (Doddington, 2005). In the EDEN project this group of accepted ARTIFACT-ARGs was expanded to include all goods (*mat* 'food', *klær* 'clothes', and *aksjer* 'stocks') that can be owned, including resources (*e.g.*, *fornybar energi* 'renewable energy') and permits (e.g., *CO2-kvoter* 'CO2 quotas'). Transfers of services (*e.g.*, *flybilletter* 'flight tickets') will not trigger TRANSFER-OWNERSHIP events.

Adaptation to the Norwegian society The ACE guidelines were constructed for annotation of English, mostly US-based news sources (Doddington et al., 2004). In the application of these guidelines to Norwegian news text, several adaptations to Norwegian language and society were deemed necessary. The event types ELECT, START-POSITION, and END-POSITION were adapted to better capture the Norwegian political scene by expanding the PERSON-ARG role of the events to also cover organizations (ORG) such as political parties (*e.g.*, the political party *Høyre* in *Høyre vant valget*, 'Høyre won the election'). The fact that the change of role that a political party goes through as a result of elections usually also results in role changes for the individual politicians supports our adaptation. Compared to the American political system, which ACE was developed to annotate events from, it is far less normal to describe an individual politician's change of role in the Norwegian media. A PERSON role that can only be filled by an entity denoting an individual would not be sufficient in capturing important political events, such as changes of government. Allowing for ORGs to take on the PERSON-ARG role of these events allows us to annotate a significant number of events and provides valuable information from the data.

Particle verbs as triggers While the ACE-guidelines only allow for particles to be part of triggers when directly adjacent to the main verb, we enriched the annotation guidelines to support the annotation of Norwegian particle verbs as triggers (*e.g.*, *støte på*, 'run into' which can trigger MEET-events), also when the particles are separated from the main verb by one or more tokens. Particle verbs, through the presence of their particle, gain semantic meaning separate from that of the verb alone (*støte* by itself means 'thrust'). *Støte* ***på*** can trigger a MEET-event, but *støte* can not. Verbal triggers are only expanded to include a particle when the verb alone does not trigger the event. If the particle is not immediately following the verb (*e.g.*, **trenge inn** as trigger for TRANSPORT in *I natt* **trengte** *store grupper migranter seg* **inn**, 'tonight big groups of migrants **forcefully entered**') we annotate the trigger as a discontinuous span. Any entities separating the two will not be annotated as part of the trigger.

3.3.6. Adaptation to transcribed speech

As expected, the text in the portion of the data from automatically transcribed spoken broadcasts was of more varying quality than the data in the NDT portion of the dataset. Several adaptations of the guidelines were deemed necessary to annotate the events present in the dataset.

Transcription errors One challenge in annotating events in transcribed speech is how to handle transcription errors. If the only potential trigger for an event is transcribed incorrectly (*e.g.*, *død* 'death', which would have triggered a DIE-event transcribed as *sdøe* in *I dag sørger fans over hele verden og ved dyvik husby sdøe* 'Today fans all over the world mourn Dyvik Husby sdeate') it is not possible to annotate the event. These cases were therefore left unannotated in the dataset.

Parts of phrasal entities (*i.e.* NPs), that make up event arguments, containing transcription errors, were replaced with new entities that only contain

the correct sub-parts of the original entity (*e.g.*, the entity *Casper gud*, which should have been "Casper Rud" being annotated as an argument consisting only of *Casper* for the MEET event: *Grekeren skulle etter planen **møte** Casper gud i kveld*, 'The Greek should according to plan **meet** Casper gud this evening'). If the only potential entity for an argument role is erroneously transcribed to the point where it is illegible (*e.g.*, *distanser* 'distances' as a wrongly transcribed name of the victim of the DIE event: *Han hevder at han er fullstendig uskyldig i dette, og at han ikke har noe med med **drapet** på distanser* 'He claims that he is completely innocent in this, and claims that he did not have anything to do with the **murder** of distances'), we do not annotate it, and leave the argument role empty. For event types with mandatory arguments (*e.g.*, PHONE-WRITE) we do not annotate the event if the argument roles are left empty.

Repetitions In addition to errors stemming from the transcription process, we also identified characteristic spoken language phenomena that influenced the annotation. Most notably, we found that the dataset of transcribed speech had a higher frequency of repetitions. For cases of repeated triggers (*e.g. bli* in *jeg skulle bli **bli** hovedtrener i Brann etter hvert*, 'I would become **become** main coach of Brann after a while') the last occurrence of the repetitions were chosen as trigger for the event, and the other repetitions were left unannotated. If there is a clear difference in quality between the transcriptions, we chose the best available transcription as trigger for the event (*han kpte **kjøpte** kjøøp en båt* 'he bght **bought** bot a boat').

3.4. Annotation procedure

The data annotation was performed using the Brat annotation software (Stenetorp et al., 2012). Four students with a background in NLP and linguistics worked as annotators of the corpus and received financial remuneration for their annotation work. All annotators were initially tasked with annotation of a small subset of the data, followed by a round of discussion and updates to the guidelines. Inter-annotator agreement calculations were performed on the final part of the news dataset as well as the transcribed spoken data.

3.5. Inter-annotator agreement

Agreement scores were calculated using Cohen's kappa (Cohen, 1960). Inter-annotator (IAA) scores for event identification was performed at the token level, while trigger label and argument label IAA scores were calculated given that both annotators agreed on an event span. Table 2 presents the IAA scores for the EDEN dataset.

Dataset	Event trigger κ	Label κ	Arg κ
TV 2	0.95	0.99	0.94
NDT	0.80	0.91	0.83
Total	0.83	0.93	0.86

Table 2: Inter-annotator agreement scores for the two final IAA datasets.

We find that annotators largely agree on event placement, but that there are some scope variations. Event type labeling shows a very high agreement, with κ=0.93, and was calculated given that the annotators agreed on the scope of the event trigger. Only 28 events were disagreed upon, with the least agreed upon labels being ATTACK (8 instances) which were often confused with INJURE. Further, for the label END-ORG (6 instances), we found disagreements with ATTACK and START-ORG. The annotators also disagreed somewhat on TRANSPORT, TRANSFER-OWNERSHIP, and TRANSFER-MONEY events. Annotators generally show a high level of agreement for event arguments with κ=0.86, and as the annotation tools prevent annotators from adding arguments not associated with a certain event, all disagreements stem from one annotator not identifying one or more arguments that the other annotator has added. The most common being PLACE-ARG and TIME-ARG.

Pre-annotations seem to have been effective for argument span identification, and as many as 69.5% of arguments were chosen from pre-annotated spans.

4. Dataset statistics

Table 3 shows the main statistics of the EDEN dataset, in terms of number of documents, sentences, tokens, as well as annotated events, arguments, and attributes. We break down the statistics to also show how the distribution of news articles vs transcribed news broadcasts are distributed in the three splits train, dev, and test. For the NDT data we followed the original data splits (Solberg et al., 2014; Øvrelid and Hohle, 2016). For the TV2 data we split the data chronologically, and follow the distribution of NDT with an 80%, 10%, 10% distribution over train, dev, and test splits respectively. While EDEN is of a rather small size in terms of total number of tokens, it is in fact larger than the most used English dataset ACE (with 303,000 tokens) and comparable to Light ERE and Rich ERE (respectively 505,837 and 180,040 tokens).

As discussed in Section 3, EDEN is annotated for 33 event types and their respective arguments.

	#Docs	#Sent	#Tokens	#Events	#Arguments	#Attributes
NDT Train	273	12,916	197,540	2,476	4,012	1,416
TV 2 Train	234	8,052	164,605	2,108	3,404	1,155
Total Train	*507*	*20.968*	*362.145*	*4.584*	*7.416*	*2.571*
NDT Dev	30	1,155	19,641	206	359	133
TV 2 Dev	21	764	16,027	181	267	109
Total Dev	*51*	*1.919*	*35.668*	*387*	*626*	*242*
NDT Test	33	1,981	29,431	302	565	170
TV 2 Test	39	1,384	27,982	532	692	234
Total Test	*72*	*3.365*	*57.413*	*834*	*1.257*	*404*
EDEN in total	630	26,252	455,226	5,805	9,299	3,217

Table 3: Statistics of the EDEN dataset in terms of total number of documents, sentences, tokens, events (total number of occurrences of event types), arguments, and attributes in both the NDT and the TV 2 datasets.

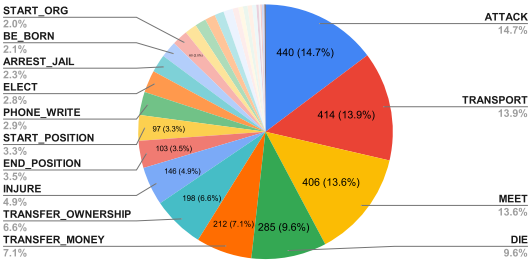

Figure 4: Event types and their occurrences in the NDT data (all splits).

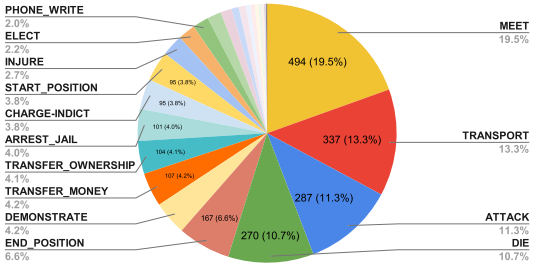

Figure 5: Event types and their occurrences in the TV 2 data (all splits).

However, the unique number of event types and event arguments annotated in each dataset and split varies. In the NDT dataset, only 29 out of the 33 event types actually occur in the training data, 23 in the dev portion of the data, and 21 in the test split. Similar trends can be seen for the transcribed data from TV 2, where 30 event types occur in train, 16 in dev, and 24 in the test split. From the original ACE event types, the two events EXTRADITE and MERGE-ORG do not occur in neither the NDT nor the TV 2 training splits. In addition to these, the event types SUE and BE-BORN were not found in the training data of NDT, and the event type PARDON was not found in the training data of TV 2.

Figure 4 and Figure 5 show the most frequent event types in each of the NDT and TV 2 datasets respectively. As evident from the statistical representations, the majority of events occur with limited frequencies, with 26 events individually occurring less than 10% in both datasets. A notable difference between both parts of the dataset is that the most frequently annotated event in NDT corresponds to the event type ATTACK, whereas in TV 2, the event type MEET has the highest frequency. Despite this difference, the top 14 most frequently occurring events in NDT are encompassed within the top 16 events in TV 2. This observation suggests a pattern in the distribution of event types within the Norwegian news domain. The data imbalance between event types is not particular to the EDEN dataset and is a known issue in for example English event datasets (Wang et al., 2020).

5. Experiments

We here present a set of experiments aiming to benchmark the EDEN dataset as a dataset for training of Norwegian event extraction systems.

5.1. Event extraction model

Our model is adapted from JSEEGraph (You et al., 2023), a graph-based model for joint structured event extraction. In a similar convention, we transform each sentence into a graph representation,

	Trg-I			Trg-C			Arg-I			Arg-C		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
NDT	57.5	68.0	62.3	56.3	66.6	61.0	51.8	50.9	51.3	50.8	49.4	50.1
TV2	77.1	75.1	76.1	76.2	74.2	75.2	55.6	51.9	53.7	54.6	50.2	52.3
EDEN	66.7	71.7	69.1	65.6	70.5	68.0	53.5	51.3	52.4	52.5	49.8	51.5

Table 4: Experimental results on EDEN: Precision (P), Recall (R), and F1 scores. "I" corresponds to "Identification", and "C" corresponds to "Classification".

with event triggers and arguments as nodes. As mentioned in Section 3.3, some event triggers and arguments consist of disjoint text spans, and such nodes will be anchored to several spans. The model we train uses XLM-R (Conneau et al., 2020) to obtain input sequence representations, and performs event extraction via predicting nodes (trigger/argument) and constructing edges (event type-/argument role) between nodes.The model contains the following modules: 1) *Sentence representation*: we use XLM-R (Conneau et al., 2020) to obtain the contextualized embeddings of the input tokens, and further map each contextual embedding to queries via a linear layer; 2) *Node prediction*: a linear classifier classifies each query into a node, and a deep biaffine classifier anchors each node to surface tokens by biaffine attention between the contextual embeddings and queries; 3) *Edge prediction*: two deep biaffine classifiers predict edge presence between nodes and the corresponding edge label.

5.2. Evaluation metrics

We report precision (P), recall (R), and F1 scores for the following metrics:

- **Trigger**: an event trigger is correctly identified (Trg-I) if its offsets match a reference trigger, and correctly classified (Trg-C) if the event type also matches a reference trigger.
- **Argument**: an event argument is correctly identified (Arg-I) if its offsets and event type match a reference argument, and correctly classified (Arg-C) if its argument role also matches that of a reference argument.

5.3. Results and discussion

The results on the EDEN test set are shown in Table 4. The overall results are comparable to the state-of-art results on datasets annotated under the ACE guidelines (You et al., 2023). Trigger extraction scores are slightly lower, due to a more complex annotation of triggers in EDEN. More concretely, EDEN commonly contains disjoint trigger spans, while triggers in other ACE datasets are always annotated as continuous text spans. In general, we also observe a small gap between identification and classification scores for both event triggers and arguments.

In terms of text types in the dataset, trigger extraction appears to be harder on the NDT data, with considerably lower scores than those of the TV 2 data. As shown in Figure 4 and Figure 5, both NDT and TV 2 data have diverse event types, and the dominant event types are similar. In this case, genres play an important role in trigger extraction, and it is more difficult to extract event triggers from newspaper articles than from the news transcripts. A possible explanation for this observation might be the length and event density of the sentences/utterances in these text types. Furthermore, since argument extraction results are affected by trigger extraction, the model with no surprise also performs better on TV 2 data.

6. Conclusion

This paper introduces a novel event identification dataset for Norwegian, EDEN, that combines data from diverse sources and genres. This is the first data for event extraction for the Norwegian language. The dataset encompasses texts derived from conventional news outlets, alongside transcribed speech texts sourced from a national news broadcasting company. The annotations in this dataset are grounded in the well-established English dataset ACE, and we provide a comprehensive account of the annotation methodology. Furthermore, the paper elucidates the adaptation of the ACE annotation guidelines to accommodate the distinctive characteristics of the Norwegian language and the source materials.

We also present our assessments of the inter-annotator agreement, and discuss which event types exhibited the highest levels of disagreement among annotators. Additionally, we report results on our initial experiments using a graph-based event detection model, establishing a benchmark model tailored to this dataset.

EDEN will be made freely available and is envisaged to serve as a foundation for future explorations into the joint extraction of event modalities along event types and associated arguments.

7. Acknowledgements

We want to thank the annotators Marie Emerentze Fleisje, Jeanett Murstad, and Lilja Charlotte Storset for their great annotation efforts.
This work was supported by industry partners and the Research Council of Norway with funding to *MediaFutures: Research Centre for Responsible Media Technology and Innovation*, through the centers for Research-based Innovation scheme, project number 309339.

8. Bibliographical References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

George Doddington. 2005. The Automatic Content Extraction (ACE) program.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Degree: A data-efficient generative event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Xiao Liu, He-Yan Huang, Ge Shi, and Bo Wang. 2022. Dynamic prefix-tuning for generative

template-based event extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5216–5228.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

David Samuel and Milan Straka. 2020. ÚFAL at MRP 2020: Permutation-invariant semantic parsing in PERIN. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789.

Huiling You, Lilja Øvrelid, and Samia Touileb. 2023. JSEEGraph: Joint structured event extraction as graph parsing. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 115–127, Toronto, Canada. Association for Computational Linguistics.

Huiling You, David Samuel, Samia Touileb, and Lilja Øvrelid. 2022. EventGraph: Event extraction as semantic graph parsing. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 7–15, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

9. Language Resource References

Doddington, George R and Mitchell, Alexis and Przybocki, Mark A and Ramshaw, Lance A and Strassel, Stephanie M and Weischedel, Ralph M. 2004. *The Automatic Content Extraction (ACE) program-tasks, data, and evaluation.* Lisbon. distributed via LDC.

Jørgensen, Fredrik and Aasmoe, Tobias and Ruud Husevåg, Anne-Stine and Øvrelid, Lilja and Velldal, Erik. 2020. *NorNE: Annotating Named Entities for Norwegian*. European Language Resources Association.

Mæhlum, Petter and Haug, Dag and Jørgensen, Tollef and Kåsen, Andre and Nøklestad, Anders and Rønningstad, Egil and Solberg, Per Erik and Velldal, Erik and Øvrelid, Lilja. 2022. *NARC – Norwegian Anaphora Resolution Corpus*. Association for Computational Linguistics.

Øvrelid, Lilja and Hohle, Petter. 2016. *Universal Dependencies for Norwegian*. European Language Resources Association (ELRA).

Pouran Ben Veyseh, Amir and Nguyen, Minh Van and Dernoncourt, Franck and Nguyen, Thien. 2022. *MINION: a Large-Scale and Diverse Dataset for Multilingual Event Detection*. Association for Computational Linguistics.

Solberg, Per Erik and Skjærholt, Arne and Øvrelid, Lilja and Hagen, Kristin and Johannessen, Janne Bondi. 2014. *The Norwegian Dependency Treebank*. European Language Resources Association (ELRA).

Song, Zhiyi and Bies, Ann and Strassel, Stephanie and Riese, Tom and Mott, Justin and Ellis, Joe and Wright, Jonathan and Kulick, Seth and Ryant, Neville and Ma, Xiaoyi. 2015. *From light to rich ere: annotation of entities, relations, and events*.

Wang, Xiaozhi and Wang, Ziqi and Han, Xu and Jiang, Wangyi and Han, Rong and Liu, Zhiyuan and Li, Juanzi and Li, Peng and Lin, Yankai and Zhou, Jie. 2020. *MAVEN: A Massive General Domain Event Detection Dataset*. Association for Computational Linguistics.