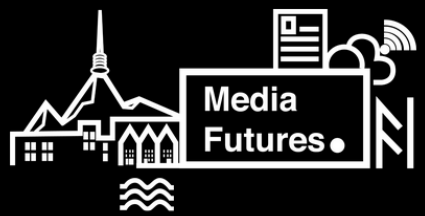
EDEN: A Dataset for Event Detection in Norwegian News

Samia Touileb, Jeanett Murstad, Petter Mæhlum,
Lubos Steskal, Lilja Charlotte Storset, Huiling You, Lilja Øvrelid

UNIVERSITETET I BERGEN

UNIVERSITETET I OSLO



EDEN – Event DEtection for Norwegian

- The first event-annotated dataset for Norwegian.

EDEN – Event DEtection for Norwegian

- The first event-annotated dataset for Norwegian.
- Focus on the news domain:
  - edited news articles,
  - transcribed spoken news broadcasts

EDEN – Event DEtection for Norwegian

- The first event-annotated dataset for Norwegian.
- Focus on the news domain:
  - edited news articles,
  - transcribed spoken news broadcasts
- We adapt ACE (Automatic Content Extraction) guidelines (Doddington, 2005) to the Norwegian news domain.

EDEN – Event DEtection for Norwegian

- The first event-annotated dataset for Norwegian.
- Focus on the news domain:
  - edited news articles,
  - transcribed spoken news broadcasts
- We adapt ACE (Automatic Content Extraction) guidelines (Doddington, 2005) to the Norwegian news domain.
- EDEN contains a total of 630 documents, annotated for 5,805 events and 9,299 arguments.

EDEN – data sources

News Text

TV News Transcripts

EDEN – data sources

News Text

News portion of the *Norwegian Dependency Treebank* (NDT)((Solberg et al., 2014; Øvrelid and Hohle, 2016))

TV News Transcripts

EDEN – data sources

News Text

News portion of the *Norwegian Dependency Treebank* (NDT)((Solberg et al., 2014; Øvrelid and Hohle, 2016))

Contains morphosyntactic annotation, named entities (Jørgensen et al., 2020), and co-reference information (Mæhlum et al., 2022).

TV News Transcripts

EDEN – data sources

News Text

News portion of the *Norwegian Dependency Treebank* (NDT)((Solberg et al., 2014; Øvrelid and Hohle, 2016))

Contains morphosyntactic annotation, named entities (Jørgensen et al., 2020), and co-reference information (Mæhlum et al., 2022).

TV News Transcripts

Norwegian television news channel *TV 2 Nyhetskanalen* provided by the broadcaster.

EDEN – data sources

News Text

News portion of the *Norwegian Dependency Treebank* (NDT)((Solberg et al., 2014; Øvrelid and Hohle, 2016))

Contains morphosyntactic annotation, named entities (Jørgensen et al., 2020), and co-reference information (Mæhlum et al., 2022).

TV News Transcripts

Norwegian television news channel *TV 2 Nyhetskanalen* provided by the broadcaster.

Transcripts produced between 2021 and 2023, but focused on a single broadcast per day (8:00 PM).
Manual inspection and rectification.

Pre-annotations

To reduce disagreement on the precise delimitation of argument entity spans.

- Use a set of heuristics over parts-of-speech and dependency relations.

Pre-annotations

To reduce disagreement on the precise delimitation of argument entity spans.

- Use a set of heuristics over parts-of-speech and dependency relations.
- Interested in locating noun phrases that are potential event arguments.

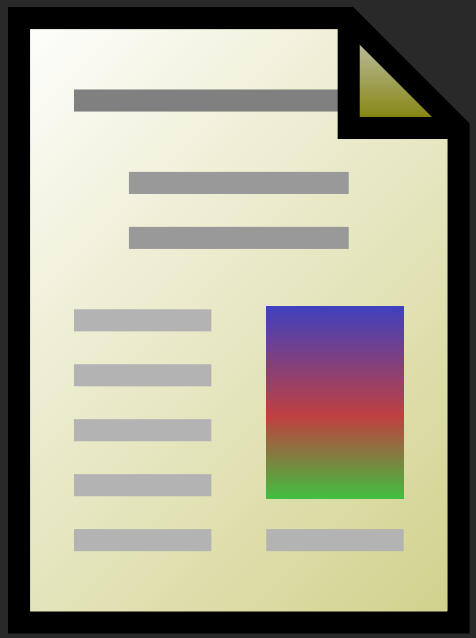
Pre-annotations

To reduce disagreement on the precise delimitation of argument entity spans.

- Use a set of heuristics over parts-of-speech and dependency relations.
- Interested in locating noun phrases that are potential event arguments.
- Pre-annotated markables are only suggestions.

Annotation guidelines

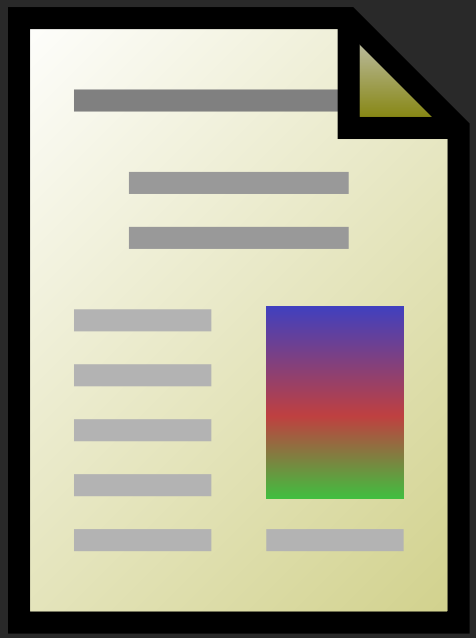
English ACE guidelines

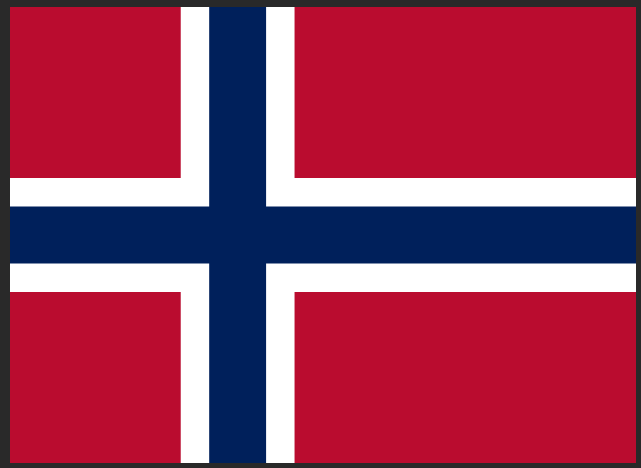


Annotation guidelines

English ACE guidelines

Adaptation





Event types

LIFE

CONTACT

JUSTICE

TRANSACTION

MOVEMENT

CONFLICT

PERSONNEL

BUSINESS

Event types

LIFE
BE-BORN
MARRY
DIVORCE
INJURE
DIE

CONTACT

JUSTICE

TRANSACTION

MOVEMENT

CONFLICT

PERSONNEL

BUSINESS

Event types

LIFE
BE-BORN
MARRY
DIVORCE
INJURE
DIE

MOVEMENT
TRANSPORT

CONFLICT

BUSINESS

CONTACT

TRANSACTION

PERSONNEL

JUSTICE

Event types

LIFE
BE-BORN
MARRY
DIVORCE
INJURE
DIE

MOVEMENT
TRANSPORT

CONFLICT
ATTACK
DEMONSTRATE

BUSINESS

CONTACT

TRANSACTION

PERSONNEL

JUSTICE

Event types

LIFE
BE-BORN
MARRY
DIVORCE
INJURE
DIE

MOVEMENT
TRANSPORT

CONFLICT
ATTACK
DEMONSTRATE

BUSINESS
START-ORG
MERGE-ORG

CONTACT

TRANSACTION

PERSONNEL

JUSTICE

Event types

LIFE
BE-BORN
MARRY
DIVORCE
INJURE
DIE

MOVEMENT
TRANSPORT

CONFLICT
ATTACK
DEMONSTRATE

BUSINESS
START-ORG
MERGE-ORG

CONTACT
MEET
PHONE-WRITE

TRANSACTION

PERSONNEL

JUSTICE

Event types

LIFE
BE-BORN
MARRY
DIVORCE
INJURE
DIE

MOVEMENT
TRANSPORT

CONFLICT
ATTACK
DEMONSTRATE

BUSINESS
START-ORG
MERGE-ORG

CONTACT
MEET
PHONE-WRITE

TRANSACTION
TRANSFER-MONEY
TRANSFER-OWNERSHIP
DECLARE-BANKRUPTCY
END-ORG

PERSONNEL

JUSTICE

Event types

LIFE
BE-BORN
MARRY
DIVORCE
INJURE
DIE

MOVEMENT
TRANSPORT

CONFLICT
ATTACK
DEMONSTRATE

BUSINESS
START-ORG
MERGE-ORG

CONTACT
MEET
PHONE-WRITE

TRANSACTION
TRANSFER-MONEY
TRANSFER-OWNERSHIP
DECLARE-BANKRUPTCY
END-ORG

PERSONNEL
START-POSITION
NOMINATE
ELECT
END-POSITION

JUSTICE

Event types

LIFE
- BE-BORN
- MARRY
- DIVORCE
- INJURE
- DIE

MOVEMENT
- TRANSPORT

CONFLICT
- ATTACK
- DEMONSTRATE

BUSINESS
- START-ORG
- MERGE-ORG

CONTACT
- MEET
- PHONE-WRITE

TRANSACTION
- TRANSFER-MONEY
- TRANSFER-OWNERSHIP
- DECLARE-BANKRUPTCY
- END-ORG

PERSONNEL
- START-POSITION
- NOMINATE
- ELECT
- END-POSITION

JUSTICE
- CHARGE-INDICT
- ARREST-JAIL
- SENTENCE
- PARDON
- EXECUTE
- FINE
- TRIAL-HEARING
- RELEASE-PAROLE
- APPEAL
- SUE
- CONVICT
- AQUIT
- EXTRADITE

Event triggers

- Event trigger:
  - (minimal) text span that most clearly describes an occurrence.

Event triggers

- Event trigger:
  - (minimal) text span that most clearly describes an occurrence.
  - mostly the main verb of a sentence.

Event triggers

- Event trigger:
  - (minimal) text span that most clearly describes an occurrence.
  - mostly the main verb of a sentence.

Hun **ringte** sønnen sin
'She **called** her son'

Event triggers

- Event trigger:
  - (minimal) text span that most clearly describes an occurrence.
  - mostly the main verb of a sentence.
  - can be a noun, participle, or adjective,

Hun **ringte** sønnen sin
'She **called** her son'

Event triggers

- Event trigger:
  - (minimal) text span that most clearly describes an occurrence.
  - mostly the main verb of a sentence.
  - can be a noun, participle, or adjective,

Hun **ringte** sønnen sin
'She **called** her son'

Byen ble rammet av et **angrep**
'The city was affected by an **attack**'

Krigen etterlot mange **skadde**
'**The war** left many **injured**'

Event triggers

- Event trigger:
  - (minimal) text span that most clearly describes an occurrence.
  - mostly the main verb of a sentence.
  - can be a noun, participle, or adjective,
  - multi-token triggers are allowed.

Hun **ringte** sønnen sin
'She **called** her son'

Byen ble rammet av et **angrep**
'The city was affected by an **attack**'

Krigen etterlot mange **skadde**
'**The war** left many **injured**'

Event triggers

- Event trigger:
  - (minimal) text span that most clearly describes an occurrence.
  - mostly the main verb of a sentence.
  - can be a noun, participle, or adjective,
  - multi-token triggers are allowed.
- Triggers can generate multiple events.

Hun **ringte** sønnen sin
'She **called** her son'

Byen ble rammet av et **angrep**
'The city was affected by an **attack**'

Krigen etterlot mange **skadde**
'**The war** left many **injured**'

Event arguments

- Denote participants involved in an event.
- Described by an event, and attributes present in the same sentence as the event trigger.

Event arguments

- Denote participants involved in an event.
- Described by an event, and attributes present in the same sentence as the event trigger.
- Entities of proper names, pronouns, terms referring to a person, organization, or GPE depending on the event type.

Event arguments

- Denote participants involved in an event.
- Described by an event, and attributes present in the same sentence as the event trigger.
- Entities of proper names, pronouns, terms referring to a person, organization, or GPE depending on the event type.
- Event types have a set of argument roles.

Event arguments

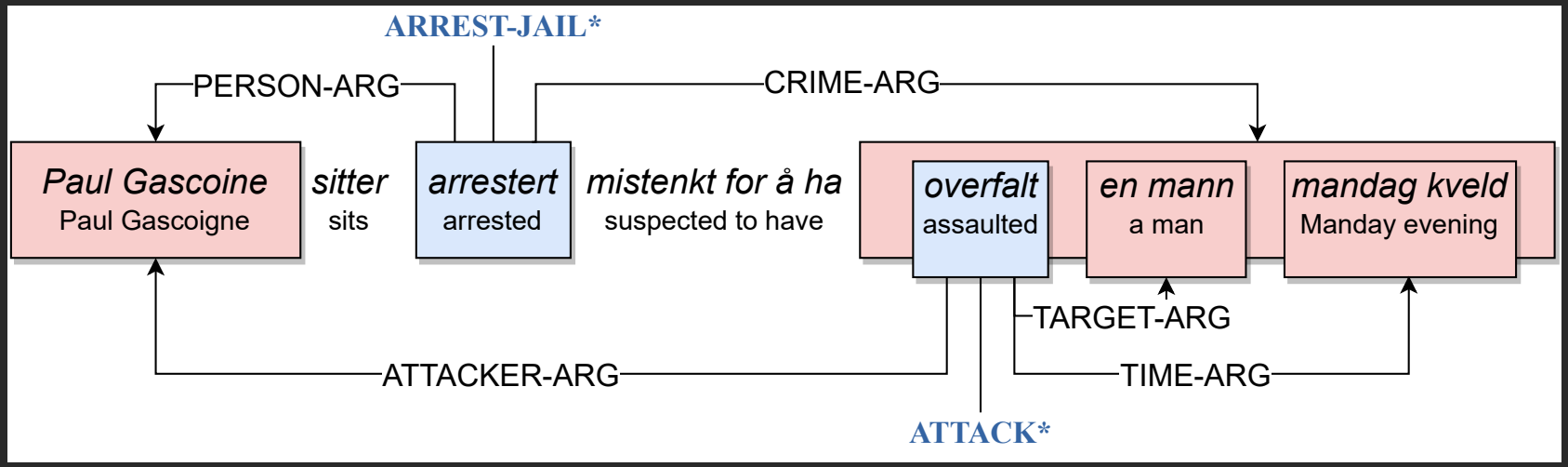


Figure: EDEN annotation showing an example of multiple events within a scope, that share an argument entity.

Event modality

- An event's modality is asserted if it is clearly described as having found place, or as currently on-going.

Event modality

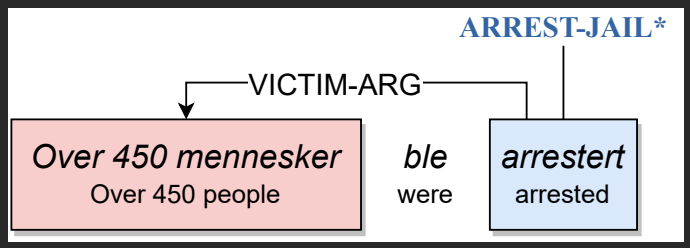


Figure: Modality asserted event annotation (indicated by asterisk).

- An event's modality is asserted if it is clearly described as having found place, or as currently on-going.
- Marked with MODAL-ASSERT.

Event modality

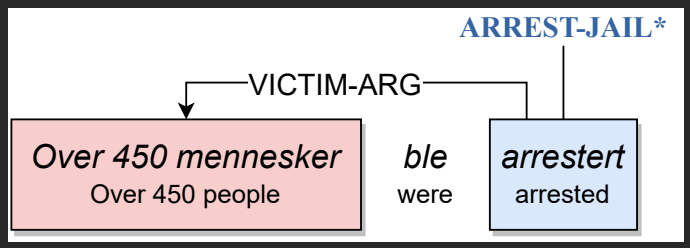


Figure: Modality asserted event annotation (indicated by asterisk).

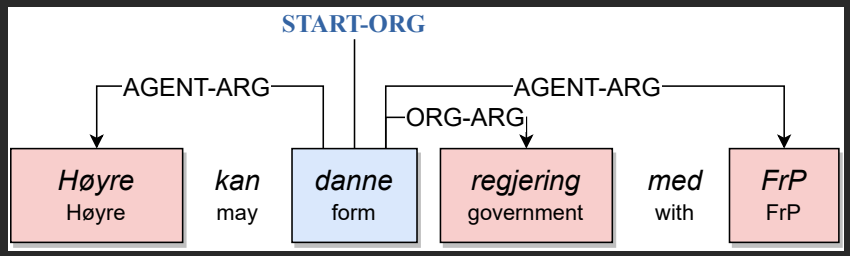


Figure: Un-asserted modality annotation.

- An event's modality is asserted if it is clearly described as having found place, or as currently on-going.
- Marked with MODAL-ASSERT.
- Un-asserted events include generic and negated events.

Adaptations of ACE guidelines

1. Clarify parts of the ACE annotation guidelines that were unclear,
2. Limit the task so as to make it manageable and provide consistent annotations, and
3. Adapt the guidelines to Norwegian language and society.

Adaptations of ACE guidelines

- **Argument selection**: entity closest to the trigger. If potential arguments are at the same distance, select the longest.

Adaptations of ACE guidelines

- **Argument selection**: entity closest to the trigger. If potential arguments are at the same distance, select the longest.
- **TIME arguments**: one overarching TIME-ARG attribute instead of the seven defined in ACE. This includes non-time expressions positioning events: ”in summer vacation”.

Adaptations of ACE guidelines

- **Argument selection**: entity closest to the trigger. If potential arguments are at the same distance, select the longest.
- **TIME arguments**: one overarching TIME-ARG attribute instead of the seven defined in ACE. This includes non-time expressions positioning events: ”in summer vacation”.
- **Transfer of ownership**: expand ARTIFACT-ARGs to include all goods (food, clothes, stocks) that can be owned, resources (renewable energy), and permits (CO2 quotas).

Adaptations of ACE guidelines

- **Argument selection**: entity closest to the trigger. If potential arguments are at the same distance, select the longest.
- **TIME arguments**: one overarching TIME-ARG attribute instead of the seven defined in ACE. This includes non-time expressions positioning events: ”in summer vacation”.
- **Transfer of ownership**: expand ARTIFACT-ARGs to include all goods (food, clothes, stocks) that can be owned, resources (renewable energy), and permits (CO2 quotas).
- **Adaptation to the Norwegian society**: adapted ELECT, START-POSITION, and END-POSITION to better capture the Norwegian political scene.

Adaptations of ACE guidelines

- **Argument selection**: entity closest to the trigger. If potential arguments are at the same distance, select the longest.
- **TIME arguments**: one overarching TIME-ARG attribute instead of the seven defined in ACE. This includes non-time expressions positioning events: ”in summer vacation”.
- **Transfer of ownership**: expand ARTIFACT-ARGs to include all goods (food, clothes, stocks) that can be owned, resources (renewable energy), and permits (CO2 quotas).
- **Adaptation to the Norwegian society**: adapted ELECT, START-POSITION, and END-POSITION to better capture the Norwegian political scene.
- **Particle verbs as triggers**: added in EDEN (*e.g.*, *støte på*, run into which can trigger MEET-events).

Adaptation to transcribed speech

- **Transcription errors**: potential triggers transcribed incorrectly are not annotated. Correct sub-parts of wrongly transcribed event arguments are annotated.

Adaptation to transcribed speech

- **Transcription errors**: potential triggers transcribed incorrectly are not annotated. Correct sub-parts of wrongly transcribed event arguments are annotated.
  - The Greek should according to plan **meet** Casper gud this evening.

Adaptation to transcribed speech

- **Transcription errors**: potential triggers transcribed incorrectly are not annotated. Correct sub-parts of wrongly transcribed event arguments are annotated.
  - The Greek should according to plan **meet** Casper gud this evening. $\rightarrow$ Casper gud should be Casper Rud, only Casper is annotated.

Adaptation to transcribed speech

- **Transcription errors**: potential triggers transcribed incorrectly are not annotated. Correct sub-parts of wrongly transcribed event arguments are annotated.
  - The Greek should according to plan **meet** Casper gud this evening. $\rightarrow$ Casper gud should be Casper Rud, only Casper is annotated.
- **Repetitions**: characteristic spoken language phenomena that influenced the annotation.

Adaptation to transcribed speech

- **Transcription errors**: potential triggers transcribed incorrectly are not annotated. Correct sub-parts of wrongly transcribed event arguments are annotated.
  - The Greek should according to plan **meet** Casper gud this evening. $\rightarrow$ Casper gud should be Casper Rud, only Casper is annotated.
- **Repetitions**: characteristic spoken language phenomena that influenced the annotation.
  - I would become **become** main coach of Brann after a while.

Adaptation to transcribed speech

- **Transcription errors**: potential triggers transcribed incorrectly are not annotated. Correct sub-parts of wrongly transcribed event arguments are annotated.
  - The Greek should according to plan **meet** Casper gud this evening. $\rightarrow$ Casper gud should be Casper Rud, only Casper is annotated.
- **Repetitions**: characteristic spoken language phenomena that influenced the annotation.
  - I would become **become** main coach of Brann after a while.
  - he bght **bought** bot a boat.

Annotation procedure

- Brat annotation software (Stenetorp et al., 2012).

Annotation procedure

- Brat annotation software (Stenetorp et al., 2012).
- Four students with a background in NLP and linguistics.
- Annotations in four phases:
  1. annotation of a small subset of the data,
  2. round of discussion and updates to the guidelines,
  3. individual annotations,
  4. overlapping annotations of the last subset of EDEN.
- Inter-annotator agreement calculations on the final part.

Inter-annotator agreement

Inter-annotator 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.

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

Dataset statistics

	#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: 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.

Data statistics

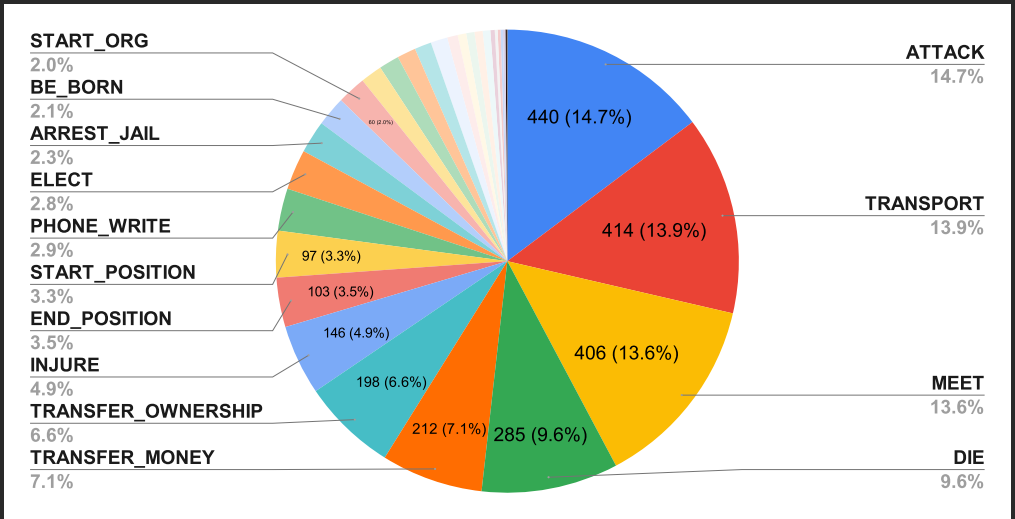


Figure: Event types in the NDT data (all splits).

Data statistics

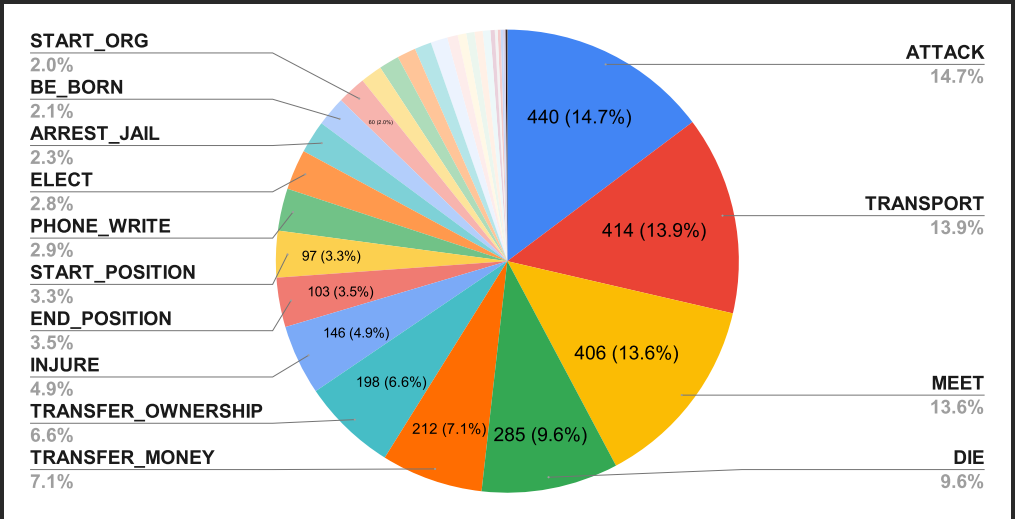


Figure: Event types in the NDT data (all splits).

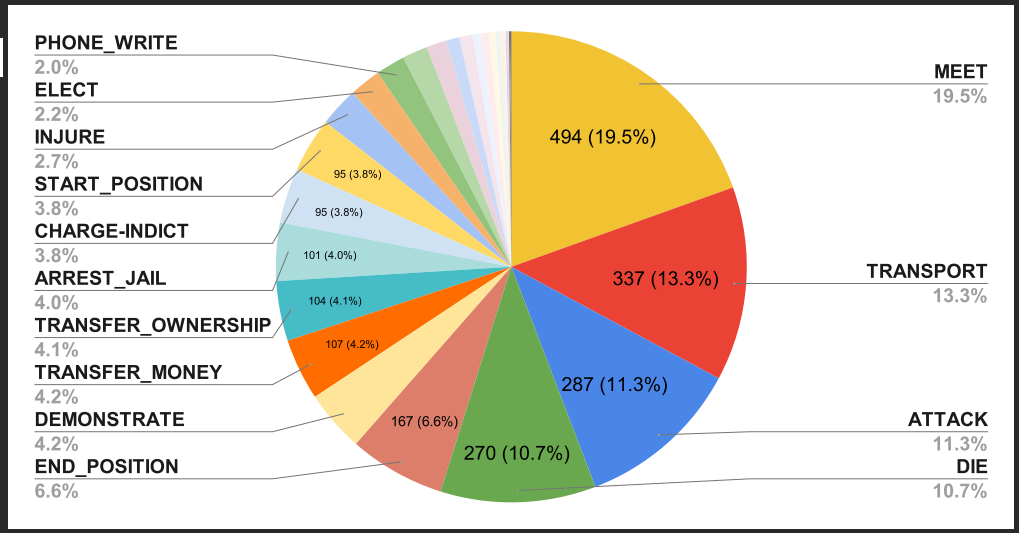


Figure: Event types and their occurrences in the TV 2 data (all splits).

Experiments

Model adapted from JSEEGraph (You et al., 2023), a graph-based model for joint structured event extraction. Each sentence is transformed into a graph representation, with event triggers and arguments as nodes.

Experiments

Model adapted from JSEEGraph (You et al., 2023), a graph-based model for joint structured event extraction. Each sentence is transformed into a graph representation, with event triggers and arguments as nodes.

- **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.

Experiments

	Trg-I	Trg-C	Arg-I	Arg-C
	F1	**F1**	**F1**	**F1**
NDT	62.3	61.0	51.3	50.1
TV2	76.1	75.2	53.7	52.3
EDEN	69.1	68.0	52.4	51.5

Table: Experimental results on EDEN. "I" corresponds to "Identification", and "C" corresponds to "Classification".

Conclusion

- EDEN: first data for event extraction for the Norwegian language.
- Conventional news outlets and transcribed speech from a national news broadcasting company.
- Annotations grounded in ACE, alongside rigorous adaptations to the Norwegian language and society.
- EDEN is freely available on our github[1].

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