Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5968–5977
Marseille, 11–16 May 2020

A New Resource for German Causal Language

Ines Rehbein and Josef Ruppenhofer
Data and Web Science Group — Archive for Spoken German
University of Mannheim, Leibniz Institute for the German Language
Mannheim, Germany — Mannheim, Germany
ines@informatik.uni-mannheim.de — ruppenhofer@ids-mannheim.de

Abstract

We present a new resource for German causal language, with annotations in context for verbs, nouns and adpositions. Our dataset includes 4,390 annotated instances for more than 150 different triggers. The annotation scheme distinguishes three different types of causal events (CONSEQUENCE, MOTIVATION, PURPOSE). We also provide annotations for semantic roles, i.e. of the cause and effect for the causal event as well as the actor and affected party, if present. In the paper, we present inter-annotator agreement scores for our dataset and discuss problems for annotating causal language. Finally, we present experiments where we frame causal annotation as a sequence labelling problem and report baseline results for the prediciton of causal arguments and for predicting different types of causation.

Keywords: Annotation of causal language, causal tagger, German

1. Introduction

Understanding causality is crucial for making sense of the world around us. Thus, understanding causal expressions in text is an important prerequisite for Natural Language Understanding (NLU). Causality, however, is also known to be a concept that defies an unambiguous definition, thus posing a challenge not only for automatic NLU systems but also for human annotators.

Several proposals have been made that describe causality from a philosophical point of view, such as the Counterfactual Theory of causation (Lewis, 1973), theories of probabilistic causation (Suppes, 1970; Pearl, 1988) and production theories like the Dynamic Force Model (Talmy, 1988). *Counterfactual Theory* tries to explain causality between two events C and E in terms of conditionals such as "If C had not occurred, E would not have occurred". However, psychological studies have shown that this not always coincides with how humans understand and draw causal inferences (Byrne, 2005). *Probabilistic Dependency theories*, on the other hand, try to explain causality based on the underlying probability of an event to take place in the world. The theory that has had the highest impact on linguistic annotation of causality is probably Talmy's *Dynamic Force Model* which provides a framework that tries to distinguish weak and strong causal forces, and captures different types of causality such as "letting", "hindering", "helping" or "intending".

While each of these theories manages to explain some aspects of causality, none of them seems to provide a completely satisfying account of the phenomenon under consideration. This problem of capturing and specifying the concept of causality is also reflected in linguistic annotation efforts. Human annotators often show only a moderate or even poor agreement when annotating causal phenomena (Grivaz, 2010; Gastel et al., 2011), or abstain from reporting inter-annotator agreement at all.

A notable exception is the work of Dunietz et al. (2015; Dunietz et al. (2017b) who, inspired by the theory of construction grammar (Goldberg, 1995), aim at building a constructicon for English causal language. When annotating these pre-defined contructions in text, Dunietz et al. (2015) obtain high agreement scores for human annotation. In the paper, we adapt their approach and present a new dataset for German causal language, with annotations in context for verbs, nouns and adpositions.

The remainder of the paper is structured as follows. First, we review related work on annotating causal language (section 2.). In section 3., we present the annotation scheme we use in the paper. Section 4. describes the annotation process and discusses problems for annotating causal language. We present baseline results for a causal tagger on our new dataset in Section 5. and end with conclusions and suggestions for future work.

2. Related Work

In this section, we give an overview over previous work on annotating causal relations.

Temporal and causal relations It has often been noted that the concept of causality is closely linked to temporal relations, as a causal relation requires the temporal order of the two events involved, and many studies have looked at both phenomena together. Mirza et al. (2014) have annotated causal relations in the TempEval-3 corpus (UzZaman et al., 2013), with an annotation scheme inspired by TimeML (Pustejovsky et al., 2010). Based on Talmy (1988) and Wolff et al. (2005), they also distinguish whether the first event causes, enables or prevents the second event. Their annotations cover different parts of speech such as verbs, adpositions, adverbials and discourse connectives.

In follow-up work (Mirza and Tonelli, 2016) they present a sieve-based system that jointly predicts temporal and causal relations in the TempEval-3 data and the TimeBank corpus (Pustejovsky et al., 2003). Their system makes use of a rich feature set, including morpho-syntactic information, syntactic dependencies, event order, WordNet similarity as well as the annotations that exist in the TimeBank corpus such as TIMEX3 attribute types or temporal signals.

Implicit vs. explicit causality It is well known that the description of causal events is not always expressed by

means of an explicit causal trigger in the text, and humans have no problem interpreting even implicit causal relations. This is exemplified in the *Causality-by-default* hypothesis (Sanders, 2005) that has shown that humans, when presented with two consecutive sentences expressing a relation that is ambiguous between a causal and an additive reading, tend to interpret the relation as causal, as in (1).

(1) She went to the pub last night. This morning, she was late for work.

The annotations in the Penn Discourse treebank (Prasad et al., 2008; Prasad et al., 2018) accomodate this phenomenon by the use of implicit discourse relations where the missing trigger is inserted by the annotators. Other work chooses to restrict themselves to annotating *causal language*, i.e. to those relations that are explicitly expressed in the text (Dunietz et al., 2015; Mirza et al., 2014). We follow the latter approach and only consider causal events that are grounded in lexical expressions in the text, ignoring implicit causal relations such as in (1) above.

Bootstrapping causal relations Many studies have tried to bootstrap causal relations, based on external knowledge bases (Kaplan and Berry-Rogghe, 1991; Girju, 2003) or on parallel or comparable corpora (Versley, 2010; Hidey and McKeown, 2016; Rehbein and Ruppenhofer, 2017).

Girju (2003) has tried to detect instances of noun-verb-noun causal relations in WordNet glosses, such as *starvation*$_{N_1}$ *causes bonyness*$_{N_2}$. After identifying noun pairs that might express a causal relation, she uses the extracted pairs to search for verbs in a large corpus that might link the nouns and express the causal relation. She then collects these verbs and obtains a list of ambiguous verbs that might express causality. To disambiguate them, Girju extracts sentences from a large text corpus and manually annotates them, according to whether or not they have a causal meaning. The annotated data is then used to train a decision tree classifier.

In previous work, we adapt this approach to a multilingual setup where we use the English verb *cause* as a seed to identify transitive causal verbs (Rehbein and Ruppenhofer, 2017). In contrast to Girju's monolingual WordNet-based approach, we use a parallel corpus and project the English tokens to their German counterparts, in order to extract and annotate causal verbal expressions for German. The extracted verbs are included in our corpus and are now augmented by additional annotations of causal nouns and adpositions.

A similar, but monolingual approach was taken by Hidey and McKeown (2016) who use two comparable English corpora, English Wikipedia and simple Wikipedia, to bootstrap causal relations. As seed data they use explicit discourse connectives from the PDTB (Prasad et al., 2008), with the aim to identify alternative lexicalisations for causal discourse relations.

Also focussing on explicit discourse relations, Versley (2010) presents a multilingual approach for data projection. He classifies German explicit discourse relations without German training data, solely based on the English annotations projected to German via word-aligned parallel text. He also presents a bootstrapping approach for a connective dictionary that relies on distribution-based heuristics on word-aligned German-English text.

Other studies on German have also been focussed on discourse connectives. Stede et al. (1998; 2002) created a lexicon for German discourse markers, augmented with semantic relations (Scheffler and Stede, 2016). Gastel et al. (2011) present annotations for discourse connectives in the TüBa-D/Z (Telljohann et al., 2004), including a small number of causal connectives. A rule-based system for detecting a set of 8 causal German discourse connectives in spoken discourse has been presented by Bögel et al. (2014). Their system predicts whether or not a connective is causal and they also try to predict the causality type, i.e. *Reason* or *Result*.

Automatic prediction of causal relations in text Dunietz et al. (2017a) present a classical feature-based system for causal tagging, trained on the annotations in the BeCause corpus. Their system uses rich syntactic and lexical information and outperforms a naive baseline.

In follow-up work, Dunietz et al. (2018) model the prediction of causal relations as a *surface construction labelling* task which can be seen as an extension of shallow semantic parsing to more complex multi-word triggers with non-contiguous argument spans. Their new system is a transition-based parser, extending the transition system of the Propbank semantic parser of Choi and Palmer (2011) for the prediction of causal constructions. The transition system is integrated in the LSTM parser of Dyer et al. (2015) which is used to compute the features for the transition system. The system operates in two steps. First, it tries to identify the causal triggers in the text, and then it labels the argument spans, i.e. cause, effect and means. The new system not only makes the time-consuming feature-engineering of earlier work superfluous, it also outperforms the previous system by a large margin.

Another neural approach for causal language detection is presented by Dasgupta et al. (2018) who extract cause-effect relations from text. They combine a bidirectional LSTM with linguistic features and use word and phrase embeddings to model the similarity between different causal arguments of the same type, e.g. the similarity between the two events 'engine failure' and 'engine breakdown'.

3. Annotation Schema

Our annotaiton scheme is adapted from Dunietz et al. (2015), but with an extended set of arguments. While Dunietz et al. (2015) annotate exactly two arguments, namely CAUSE and EFFECT, we also consider the ACTOR and the AFFECTED party of the causal event. The motivation behind our decision to extend the argument set is that we would like to add some FrameNet flavor to the PDTB-style annotations of Dunietz et al. We thus aim at providing a description of causal events and their participants, similar to FrameNet-style annotations (Ruppenhofer et al., 2006) but at a more coarse-grained level. FrameNet offers a high number of different causal frames with detailed descriptions of the actors, agents and entities involved in the

Type	Definition
CONSEQUENCE	instances assert that the Cause naturally leads to the Effect via some chain of events, without highlighting the conscious intervention of any agent. The majority of instances are Consequences.
MOTIVATION	instances assert that some agent perceives the Cause, and therefore consciously thinks, feels, or chooses something. Again, what is important for this scheme is how the relationship is presented, so an instance is Motivation only if it frames the relationship in a way that highlights an agent's decision or thought.
PURPOSE	instances assert that an agent chooses the Effect out of a desire to make the contents of the Cause span true. What distinguishes Purposes from Motivations is whether the motivating argument is a fact about the world or an outcome the agent hopes to achieve.

Table 1: Definition of the three types of causation (see Dunietz 2018, pp.33).

event.[1] For instance, FrameNet captures details such as the intentionality of the triggering force, to express whether or not the action was performed volitionally.

In our work we aim at a more generic representation that captures different types of causality, and that allows us to generalise over the different participants and thus makes it feasible to train an automatic system by abstracting away from individual lexical triggers. The advantage of such an approach is a greater generalisability and thus higher coverage. Our annotation scheme includes the following four participant roles:

1. CAUSE – a force, process, event or action that produces an effect
2. EFFECT – the result of the process, event or action
3. ACTOR – an entity that, volitionally or not, triggers the effect
4. AFFECTED – an entity that is affected by the results of the cause

We hoped that distinguishing between CAUSE and ACTOR will help the system to learn selectional preferences that some causal triggers have for specific participant roles, and also provide more informative output for applications. Compare, for instance, examples (2) and (3). The two argument slots for the verbal triggers *erzeugen* (produce) and *erleiden* (suffer) are filled with different roles. The subject slot for *erzeugen* expresses either CAUSE or ACTOR and the direct object encodes the EFFECT. For *erleiden*, on the other hand, the subject typically realises the role of the AFFECTED entity, and we often have the CAUSE or ACTOR encoded as the adpositional object of a *durch* (by) PP.

(2) **Elektromagnetische Felder**$_{Cause}$ können *Krebs*$_{Effect}$ erzeugen.
Electromagnetic fields can cancer produce.
"Electromagnetic fields can cause cancer."

(3) **Länder wie Irland**$_{Affected}$ werden durch die Reform *massive Nachteile*$_{Effect}$ erleiden.
Countries like Ireland will by the reform massive disadvantages suffer.
"Countries like Ireland are to be badly affected by the reform."

Given that there are systematic differences between prototypical properties of the participants (e.g. an ACTOR is usually animate and a sentient being), and also in the way how they combine and select their predicates, we preserve this information and see how this affects results when training an automatic system (see Section 5.).

In addition to the participants of a causal event, we follow Dunietz et al. (2015) and distinguish three different types of causation (CONSEQUENCE, MOTIVATION, PURPOSE; see Table 1), and two degrees (FACILITATE, INHIBIT). The degree distinctions are inspired by Wolff et al. (2005) who see causality as a continuum from total prevention to total entailment, and describe this continuum with three categories, namely CAUSE, ENABLE and PREVENT. Dunietz et al. (2015) further reduce this inventory to a polar distinction between a positive causal relation (e.g. *cause*) and a negative one (e.g. *prevent*), as they observed that human coders were not able to reliably apply the more fine-grained inventories.[2] The examples below illustrate the different types of causation.[3]

(4) *Auch für die Wirtschaft*$_{Affected}$ habe$_{Support}$ **dies**$_{Cause}$ bedenkliche Folgen. CONSEQUENCE
"**This**$_{Cause}$ would$_{Support}$ also have$_{Support}$ serious consequences *for the economy*$_{Affected}$."

(5) **Diese Station**$_{Actor}$ hatte *bei den ukrainischen Machthabern*$_{Affected}$ offensichtlich *Missfallen*$_{Effect}$ erregt. MOTIVATION
"**This station**$_{Actor}$ had obviously aroused *displeasure*$_{Effect}$ *among the Ukrainian rulers*$_{Affected}$."

(6) **Mehr Gewinn und mehr Arbeit**$_{Cause}$ durch *Vermeiden*$_{Effect}$ PURPOSE
" **More profit and more work**$_{Cause}$ by *avoiding*$_{Effect}$"

[1] Also see Vieu et al. (2016) for a revised and improved treatment of causality in FrameNet.

[2] For the polar distinction, they report perfect agreement.

[3] For more details on the annotation scheme, we refer the reader to Dunietz et al. (2015; 2018).

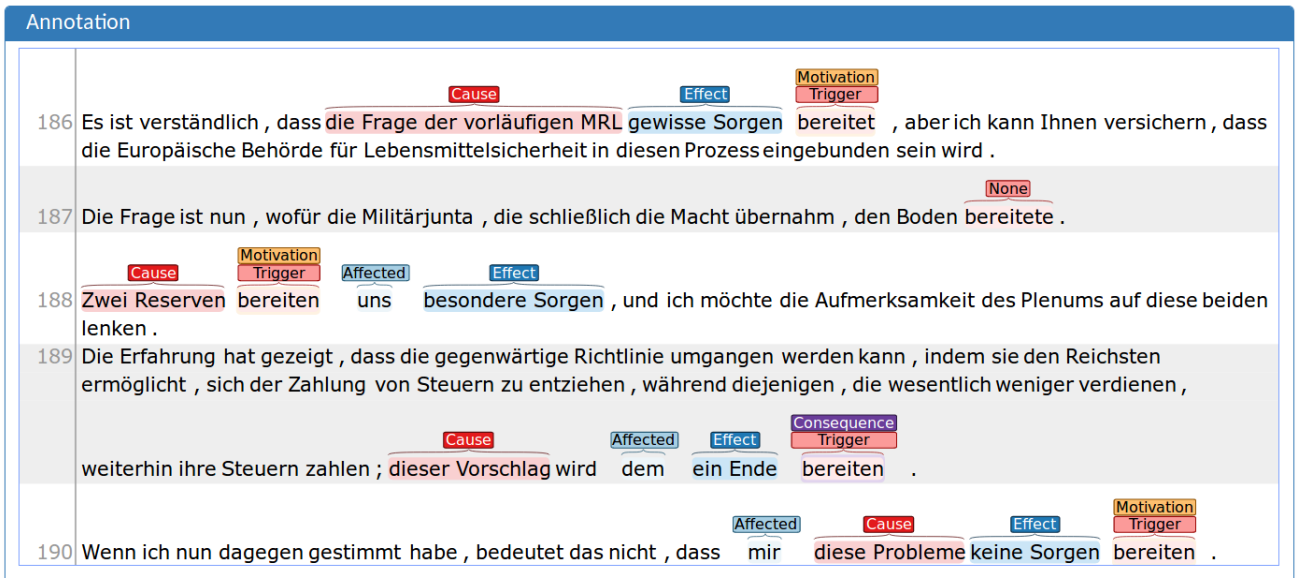

Figure 1: Causal annotations for the verb *bereiten*, visualised in the annotation tool Webanno.

While the first version of the BeCause corpus included the label INFERENCE for epistemic uses of causality, this label was given up in version 2.0 of the corpus (Dunietz et al. (2017a). We decided to follow their decision and only consider three types of causality.

4. Annotating German Causal Language

This section presents the data and annotation setup as well as inter-annotator agreement scores for the annotation of German causal language.

4.1. Data

The data we annotate comes from two sources, (i) newspaper text from the TiGer corpus (Dipper et al., 2001) and (ii) political speeches from the Europarl corpus (Koehn, 2005). We chose those two sources as we wanted to include medially written and spoken data, and we selected corpora that allow us to make the annotated data available to the research community.

4.2. Annotation Experiment

The annotation was done by three annotators, two expert annotators and one advanced student of Computation Linguistics. Each instance included only one causal trigger to be annotated. The triggers were marked and the annotators had been instructed to ignore other potentially causal expressions in the same sentence. For annotation, we used the online tool WebAnno (Yimam and Gurevych, 2013) (Figure 1). Each instance in the dataset was annotated by at least two annotators, and after the annotation process was completed, all disagreements were resolved by the two expert annotators. Table 2 shows inter-annotator agreement (IAA) scores for the different subsets of our data. We re-

Source	POS	Args (κ)	Types (κ)	# types
Europarl	*verb*	0.94	0.695	912
TiGer	*noun*	0.98	0.900	1,158
Europarl	*adp*	0.93	0.782	977
TiGer	*adp*	0.93	0.810	1,272

Table 2: IAA (Fleiss' κ) for annotation of verbs, nouns and adpositions for causal arguments and causal types.

Confusion matrix for causal types (verbs)

ANNOT 1 / ANNOT 2	CONSEQ.	MOTIV.	PURPOSE	NONE
CONSEQ.	498	14	1	3
MOTIV.	75	76	0	0
PURPOSE	7	1	11	1
NONE	41	8	0	176

Confusion matrix for causal types (adpositions)

ANNOT 1 / ANNOT 2	CONSEQ.	MOTIV.	PURPOSE	NONE
CONSEQ.	337	45	7	42
MOTIV.	54	353	1	16
PURPOSE	7	8	163	37
NONE	28	28	31	1,126

Confusion matrix for causal types (nouns)

ANNOT 1 / ANNOT 2	CONSEQ.	MOTIV.	PURPOSE	NONE
CONSEQ.	328	11	0	10
MOTIV.	37	370	0	5
PURPOSE	0	0	47	0
NONE	7	9	1	333

Figure 2: Confusion matrices for causal type annotations for verbs, nouns and adpositions.

Source	POS	# forms	# instances	% causal	Consequence		Motiv.		Purpose	
Europarl	verb	112	932	78.9	76.3%	(561)	22.0%	(162)	1.6%	(12)
TiGer	noun	21	1,178	69.3	43.9%	(359)	50.2%	(410)	5.9%	(48)
TiGer	adp	26	983	40.9	43.3%	(174)	42.3%	(170)	14.4%	(58)
EuroParl	adp	26	1,297	54.7	39.3%	(279)	36.1%	(256)	24.5%	(174)
Total		**159**	**4,390**	**60.7**	**51.5%**	**(1,373)**	**37.5%**	**(998)**	**11.0%**	**(292)**

Table 3: Corpus statistics for the annotations of the German Causal Constructicon (version 1.0).

port Fleiss' κ[4] for the different annotation layers (Table 2). As shown in Table 2, the annotation of causal arguments seems to be much easier than determining the causal type. For causal participants, we achieve IAA scores in the range of 0.93–0.98 Fleiss' κ. For the three causal types, however, agreement is much lower. While for nouns and adpositions the agreement for causal types was satisfactory (0.78–0.90 Fleiss' κ), IAA for verbs was substantially lower with a Fleiss' κ of 0.69. Below we discuss hard cases that caused disagreemets between the human annotators.

4.3. Discussion: Hard Cases

While the agreement for nouns and adpositions is fairly high, we notice many disagreements for verbs. The confusion matrix for verbs shows that here the annotators seemed to struggle between the causal types CONSEQUENCE and MOTIVATION (Figure 2).

When looking into the data, we see that one reason for the lower agreement for verbs are cases where the definition for MOTIVATION in our guidelines was not specific enough and has thus been interpreted differently be the annotators. Our definition followed the one of Dunietz (2018) below.

> "MOTIVATION instances assert that some agent perceives the Cause, and therefore consciously thinks, feels, or chooses something. Again, what is important for this scheme is how the relationship is presented, so an instance is Motivation only if it frames the relationship in a way that highlights an agent's decision or thought." Dunietz (2018, p.33)

We observe that the following verbs *auslösen, hervorrufen, stiften, erregen, bereiten* (trigger, elicit, cause, attract, cause) are frequently used in contexts where a stimulus provokes an experiencer to react emotionally or psychologically (Examples 7 a-e).

(7) a. Die Möglichkeit übermäßig komplizierter administrativer Anforderungen bereitet mir Sorge.

"The possibility of excessively complicated administrative requirements worries me."

b. Dieser Unfall versetzte die Welt in Angst und Schrecken.

"This accident frightened the world."

c. Die Spannungen haben Unbehagen in Europa hervorrufen.

"The tensions have caused unease in Europe."

d. Die Unfälle haben die Aufmerksamkeit der Union erregt.

"The accidents have attracted the attention of the Union."

e. Dieses Thema hat große Besorgnis in der Bevölkerung ausgelöst.

"This issue has caused great concern among the population."

One annotator interpreted those instances as MOTIVATION, based on the fact that they highlight an agent's (or a group of agents') state of mind. The second annotator, however, assumed that instances of MOTIVATION would require an agent's thought process that resulted in an intentional action or decision (which is not the case for the examples above) and thus annotated the same instances as CONSEQUENCE. Given that these cases are quite frequent in our dataset, they account for a large part of the disagreements between the annotators. During adjudication, we specified the guidelines and decided to consider those instances as cases of MOTIVATION.

4.4. A Corpus of German Causal Language

The new resource comprises 4,390 annotated instances with more than 5,000 annotated causal arguments for three different parts of speech (verbs, nouns and adpositions) Table 4 shows the number of annotated causal arguments in the corpus. The distribution of different types of causation in the new dataset can be seen in Table 3. The data is available in WebAnno TSV 3.2 format, a tab-separated format similar to CoNLL.[5]

POS	Actor	Affected	Cause	Effect	All
verb	96	235	625	711	1,667
noun	64	77	511	514	1,166
adp	16	0	1,087	1,100	2,203
Total	**176**	**312**	**2,223**	**2,325**	**5,036**

Table 4: Distribution of causal arguments across different parts-of-speech.

5. Experiments

In this section, we present first experiments on automatically predicting German causal language. We split the

[4] We use the R package 'irr' for computing inter-annotator agreement: https://cran.r-project.org/web/packages/irr/index.html.

[5] The data is available from https://github.com/josefkr/causal_annotations_DE.

problem into two tasks, (i) the prediction of causal arguments such as CAUSE, EFFECT, ACTOR and (ii) predicting the type of causality, i.e CONSEQUENCE, MOTIVATION or PURPOSE. We describe our training and test data and introduce our system before we present baseline results for both tasks on our new dataset.

5.1. Data and Model

Data preparation We divided the data into training, development and test set with 86,797 / 3,899 / 35,803 tokens, respectively. The distribution of (causal and non-causal) triggers in the data is shown in Table 5.[6]

	Tokens	Sent.	Trigger	causal	non-causal
train	86,797	2,915	2,937	1,787	1,150
dev	3,899	151	151	78	73
test	35,803	1,336	1,377	873	504

Table 5: German causal annotation dataset split into training/development/test sets.

Model: A BERT-based causal sequence tagger We model the task of causal language prediction as a sequence labelling problem, following related work on local semantic role labelling employing syntax-agnostic neural methods (Collobert et al., 2011). Our system is a neural sequence tagger based on Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2019).
Recently, transformers have pushed the state of the art for many NLP applications by learning context-sensitive embeddings with different optimisation strategies and then fine-tuning the pre-trained embeddings in a task-specific setup. BERT embeddings are usually trained on large amounts of data, incorporating word embeddings with positional information and self-attention. The representations are trained in two different task setups, i.e. by predicting masked words based on their left and right context and by classifying two sentences based on how probable it is that the second one immediately succeeds the first one in a text document. As a result, the learned embeddings encode information about the left and right context for each word which makes them superior to most previous representations.
Devlin et al. (2019) have proposed a BERT architecture for sequence tagging on the CoNLL-2003 NER shared task data (Sang and Meulder, 2003). The model uses the pre-trained BERT embeddings for initialisation and then fine-tunes the representations by adding a simple classification layer on top of the pre-trained BERT model and jointly fine-tuning the model parameters on the downstream task. Each BERT model provides its own tokenisation which splits longer words into sub-tokens. The sequence tagger uses only the first sub-token as the input to the classifier, which then predicts a label for each token.

[6] The mismatch between the number of sentences and triggers is caused by German particle verbs where the particle can be split from the verb stem. This results in two trigger words per sentence, as in the example for *auslösen* (trigger): *Der Brand löste eine Explosion aus.* (The fire set off an explosion).

We use the HuggingFace transformers library (Wolf et al., 2019) that provides pre-trained transformer models for different languages and tasks. The model we choose for our experiments is the pre-trained German uncased BERT model (bert-base-german-dbmdz-uncased).[7]

5.2. Prediction or Causal Arguments

For causal argument prediction, we provide the tagger with the lower-cased word forms. In each training, development and test sentence there is exactly one target token that we mark in the input so that the tagger can learn that this token is a potential trigger. An exception are sentences with split particle verbs, as in Example (8) below. Here we mark both, the verb stem and the separated particle, in the input. However, only the verb stem is tagged with a causal label in the gold standard while the verb particle is labelled as 'O' (outside of any causal argument).

(8) der Schiffbruch löste eine Welle der Empörung aus
the shipwreck triggered a wave of indignation PTCL

Please note that not all marked words have a causal meaning. The tagger has thus to determine whether or not the marked word form is causal (TRIGGER) or not (NONE).[8]
We let the tagger predict the causal arguments at the same time, and also predict the auxiliary tags SUPPORT and CONTROLLER.
These auxiliary tags are used to mark support and control predicates that are outside of the maximal projection of the causal trigger (see Examples 9 and 10). Our approach is similar to the treatment of support and control predicates in the German SALSA corpus (Rehbein et al., 2012).
For verbs, the causal arguments are almost always inside the maximal syntactic projection of the trigger. This, however, is not the case for nouns where the causal participants are not always governed by the nominal trigger (Examples 9 and 10). The support and control predicates thus allow us to annotate all causal arguments, even those that are outside the trigger's projection.

(9) Der Anlass *für den Volkszorn* war$_{Support}$ **der Tod eines Kindes**.
"The reason for the people's anger was the death of a child."

(10) Zwei Studien nennen$_{Controller}$ *für den Wählerschwund* als$_{Controller}$ Grund auch **unprofessionelle Führung**.
"Two studies cite unprofessional leadership as a reason for voter shrinkage."

This gives us the following label set for the tagger to predict: TRIGGER, CAUSE, EFFECT, ACTOR, AFFECTED,

[7] The model has been trained by the MDZ Digital Library team (dbmdz) at the Bavarian State Library on 2,350,234,427 tokens of raw text, including Wikipedia, the EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl. For details see `https://github.com/dbmdz/german-bert`.

[8] Note that NONE signifies the trigger words that have a non-causal meaning while O refers to all tokens that are neither a potential trigger word nor inside any causal argument.

ID	F1	Trigger	Cause	Effect	Actor	Affected	None	O	Support	Controller
1	72.8	91.0	83.3	83.3	4.1	65.0	88.4	94.2	77.7	68.5
2	71.5	91.0	82.6	83.2	0	62.6	88.1	94.2	76.4	65.7
3	74.4	91.8	83.8	84.0	14.6	63.1	89.3	94.4	78.8	70.1
4	70.2	90.8	82.5	83.4	0	57.8	87.4	93.8	74.1	62.0
5	72.3	91.3	83.2	83.8	0	61.9	87.3	94.3	76.5	72.7
avg.	**72.2**	**91.2**	**83.1**	**83.5**	**3.7**	**62.1**	**88.1**	**94.2**	**76.7**	**67.8**

Table 6: Results for the prediction of causal triggers and causal arguments on the new dataset.

SUPPORT, CONTROLLER, NONE, O (see footnote 8).

Table 6 shows results for each of the causal participants. Our results show that the tagger is able to learn to predict most of the tags with a reasonable performance. Exceptions are the semantic roles ACTOR and AFFECTED for which we only have a small number of training instances (176 instances of ACTOR and 312 instances of AFFECTED in the combined train/dev/test data).

The tagger is also able to distinguish between causal and non-causal uses of triggers, with an F1 of 91.2% for causal and an F1 of 88.1% for non-causal readings.

5.3. Prediction of Causal Types

Next we try to predict the causal types of our trigger words. We use the same sequence tagging architecture but this time we remove all argument labels from the training data but add the causal types for each trigger. The task of the tagger is now to decide whether a trigger word has a causal or non-causal reading in a particular context, and also to predict the type of causality for the causal triggers. This gives us five labels: CONSEQUENCE, MOTIVATION, PURPOSE, NONE, O. Again, we mark the trigger words in the input to let the tagger know which token(s) are potential causal triggers.

We report results for a most frequent sense (MFS) baseline where we map each word form to their corresponding lemma form and, for each lemma, always predict its most frequent sense in the training data (including NONE for non-causal readings). For lemmas not seen in the training data, we predict the NONE class. This gives us a strong baseline with an overall F1 of 81.17% and an F1 for the three causal labels between 72.5% and 78.9% F1 (Table 7). The high score for the PURPOSE class is due to the fact that we have some nouns and adpositions that are always causal and only invoke the PURPOSE sense (e.g. for adpositions: *halber, zwecks* (for the sake of, in order to), for nouns: *Zweck* (purpose)). This makes the MFS baseline for PURPOSE particularly hard to beat.

Table 7 shows results for the prediction of causal types on non-gold triggers, i.e. where the tagger also has to decide whether the trigger is causal or not, in addition to predicting the correct causal type. Not surprisingly, we get best results for the most frequent label (CONSEQUENCE, F1 83.22%). Results for the other two labels, MOTIVATION and PURPOSE, are also quite high with 79.87% and 78.64% F1, respectively. For PURPOSE, however, we fail to outperform the MFS baseline while for the other two classes our tagger shows improvements over the baseline of around 5% (for MOTIVATION) and 10% (for CONSEQUENCE).

We also present an experiment on gold triggers where we only mark input triggers that, in this particular context, have a causal meaning. The objective of this experiment is to find out how well the classifier performs under optimal conditions. This could be seen as an upper bound for the prediction of causal types.

ID	F1	CONSEQ.	MOTIV.	PURPOSE	NONE
MFS	81.17	72.47	74.78	**78.87**	79.75
1	86.25	83.19	80.29	78.91	88.87
2	84.69	81.63	77.72	77.33	86.78
3	87.75	85.14	82.39	82.67	88.56
4	85.72	82.81	79.57	78.38	87.84
5	85.46	83.33	79.36	75.91	88.69
avg.	**85.97**	**83.22**	**79.87**	78.64	**88.15**

Table 7: Most frequent sense (MFS) baseline and results (F1) for the prediction of causal types on causal/non-causal triggers.

Confusion matrix for causal types (MFS baseline).

ACTUAL / PREDICT	CONSEQ.	MOTIV.	PURPOSE	NONE
CONSEQ.	308	38	2	107
MOTIV.	45	212	0	37
PURPOSE	6	1	56	14
NONE	36	22	7	439

Confusion matrix for causal types (w/o gold triggers).

ACTUAL / PREDICT	CONSEQ.	MOTIV.	PURPOSE	NONE
CONSEQ.	391	27	3	34
MOTIV.	54	224	0	16
PURPOSE	8	1	58	10
NONE	32	12	9	451

Confusion matrix for causal types (with gold triggers).

ACTUAL / PREDICT	CONSEQ.	MOTIV.	PURPOSE
CONSEQ.	421	29	5
MOTIV.	62	231	1
PURPOSE	9	1	67

Figure 3: Confusion matrices for the prediction of causal types (most frequent sense (MFS) baseline, w/o and with gold trigger information).

ID	F1	CONSEQ.	MOTIVATION	PURPOSE
1	90.37	88.91	83.24	89.33
2	88.74	87.49	80.58	86.90
3	91.82	90.73	87.52	89.04
4	90.59	88.87	84.47	89.04
5	90.70	88.84	83.68	90.28
avg.	**90.44**	**88.97**	**83.90**	**88.92**

Table 8: Results (F1) for the prediction of causal types on gold causal triggers.

Results for the prediction of causal types on gold triggers (where the tagger knows which triggers are causal and only has to predict the type of causality) are shown in Table 8. As expected, results for gold triggers are substantially higher than for predicted ones. Here scores increase by 4 to 10 % F1.

Figure 3 shows confusion matrices for causal types for (i) the most frequent sense baseline, (ii) for automatically predicted causal types without and (iii) with gold trigger information. We can see that the tagger also struggles to discriminate between CONSEQUENCE and MOTIVATION senses, as did our human annotators.

6. Conclusion

We presented a new resource for German causal language, with annotations of causal events and their participants. Our annotations distinguish between cause and effect, and also annotate the actor and affected party of the events. The dataset includes 159 distinct causal triggers (nouns, verbs and adpositions) and 4,390 instances of those triggers, annotated in context. While our IAA was satisfactory, we showed that in particular the distinction between CONSEQUENCE and MOTIVATION is often hard not only for humans but also for automatic systems.

Further, we presented experiments on automatically predicting German causal language. We trained a neural sequence tagger, based on bidirectional transformers, on our data and showed that this syntax-agnostic system is well suited to learn the annotations in our dataset, given that we have enough training instances for each category.

In future work, we would like to add more annotations especially for the causal arguments ACTOR and AFFECTED in order to improve the accuracy for automatic predictions of those roles. We will make our annotations freely available and hope that the new dataset will trigger more work on annotating and predicting causal language.

7. Acknowledgements

This work was supported in part by the Leibniz Science Campus "Empirical Linguistics and Computational Modeling", funded by the Leibniz Association under grant no. SAS-2015-IDS-LWC and by the Ministry of Science, Research, and Art (MWK) of the state of Baden-Württemberg.

8. Bibliographical References

Bögel, T., Hautli-Janisz, A., Sulger, S., and Butt, M. (2014). Automatic detection of causal relations in german multilogs. In Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), pages 20–27, Gothenburg, Sweden, April. Association for Computational Linguistics.

Byrne, R. M. (2005). The rational imagination: How people create counterfactual alternatives to reality. *Behavioral and Brain Sciences*, 30:439–453.

Choi, J. D. and Palmer, M. (2011). Transition-based semantic role labeling using predicate argument clustering. In The ACL 2011 Workshop on Relational Models of Semantics, pages 37–45, Portland, Oregon, USA. Association for Computational Linguistics.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Dasgupta, T., Saha, R., Dey, L., and Naskar, A. (2018). Automatic extraction of causal relations from text using linguistically informed deep neural networks. In The 19th Annual SIGdial Meeting on Discourse and Dialogue, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dipper, S., Brants, T., Lezius, W., Plaehn, O., and Smith, G. (2001). The TIGER Treebank. In The Workshop on Treebanks and Linguistic Theories, TLT'01, pages 24–41.

Dunietz, J., Levin, L., and Carbonell, J. (2015). Annotating causal language using corpus lexicography of constructions. In Proceedings of The 9th Linguistic Annotation Workshop, LAW IX, pages 188–196, Denver, Colorado, USA, June. Association for Computational Linguistics.

Dunietz, J., Levin, L., and Carbonell, J. (2017a). Automatically tagging constructions of causation and their slot-fillers. *Transactions of the Association for Computational Linguistics*, 5:117–133.

Dunietz, J., Levin, L., and Carbonell, J. (2017b). The BECauSE Corpus 2.0: Annotating causality and overlapping relations. In The 11th Linguistic Annotation Workshop, LAW XI, pages 95–104.

Dunietz, J., Carbonell, J., and Levin, L. (2018). DeepCx: A transition-based approach for shallow semantic parsing with complex constructional triggers. In The 2018 Conference on Empirical Methods in Natural Language Processing, pages 1691–1701, Brussels, Belgium. Association for Computational Linguistics.

Dunietz, J. (2018). Annotating and Automatically Tagging Constructions of Causal Language. Ph.D. thesis, School of Computer Science, Pittsburgh, PA 15213.

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., and Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. In The 53rd An-

nual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL'15, pages 334–343, Beijing, China. Association for Computational Linguistics.

Gastel, A., Schulze, S., Versley, Y., and Hinrichs, E. (2011). Annotation of explicit and implicit discourse relations in the tüba-d/z treebank. In Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology, GSCL'11.

Girju, R. (2003). Automatic detection of causal relations for question answering. In Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, pages 76–83, Sapporo, Japan, July. Association for Computational Linguistics.

Goldberg, A. (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago University Press.

Grivaz, C. (2010). Human judgements on causation in french texts. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, may. European Language Resources Association (ELRA).

Hidey, C. and McKeown, K. (2016). Identifying causal relations using parallel wikipedia articles. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1424–1433, Berlin, Germany, August. Association for Computational Linguistics.

Kaplan, R. M. and Berry-Rogghe, G. (1991). Knowledge-based acquisition of causal relationships in text. *Knowledge Acquisition*, 3(3):317–337.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Conference Proceedings: The tenth Machine Translation Summit, AAMT'05, pages 79–86. AAMT.

Lewis, D. K. (1973). Counterfactuals. Blackwell.

Mirza, P. and Tonelli, S. (2016). CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In The 26th International Conference on Computational Linguistics: Technical Papers, COLING'16, pages 64–75, Osaka, Japan. The COLING 2016 Organizing Committee.

Mirza, P., Sprugnoli, R., Tonelli, S., and Speranza, M. (2014). Annotating causality in the TempEval-3 corpus. In The EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems. San Francisco: Morgan Kaufman.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In LREC.

Prasad, R., Webber, B., and Lee, A. (2018). Discourse annotation in the PDTB: The next generation. In The 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Ferro, D. D. L., and Lazo, M. (2003). The timebank corpus. In Corpus Linguistics, pages 647–656, 01.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Rehbein, I. and Ruppenhofer, J. (2017). Catching the Common Cause: Extraction and Annotation of Causal Relations and their Participants. In Proceedings of the 11th Linguistic Annotation Workshop, LAW XI, pages 105–114.

Rehbein, I., Ruppenhofer, J., Sporleder, C., and Pinkal, M. (2012). Adding nominal spice to salsa – frame-semantic annotation of german nouns and verbs. In Jeremy Jancsary, editor, The 11th Conference on Natural Language Processing, KONVENS-2012, pages 89 – 97. Eigenverlag ÖGAI.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Sanders, T. J. (2005). Coherence, causality and cognitive complexity in discourse. In First International Symposium on the Exploration and Modelling of Meaning, SEM-05.

Sang, E. F. T. K. and Meulder, F. D. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003, pages 142–147.

Scheffler, T. and Stede, M. (2016). Adding semantic relations to a large-coverage connective lexicon of german. In Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC.

Stede, M. and Umbach, C. (1998). DiMLex: A lexicon of discourse markers for text generation and understanding. In The 17th International Conference on Computational Linguistics, COLING'98.

Stede, M. (2002). DiMLex: A lexical approach to discourse markers. In *Exploring the Lexicon – Theory and Computation*. Alessandria (Italy):Edizioni dell'Orso.

Suppes, P. (1970). A Probabilistic Theory of Causality. Amsterdam: North-Holland Publishing Company.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, 12(1):49–100.

Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The Tüba-D/Z Treebank: Annotating German with a Context-Free Backbone. In The Fourth International Conference on Language Resources and Evaluation, LREC'04, pages 2229–2235.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verha-

gen, M., and Pustejovsky, J. (2013). SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Versley, Y. (2010). Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In Workshop on the Annotation and Exploitation of Parallel Corpora, AEPC.

Vieu, L., Muller, P., Candito, M., and Djemaa, M. (2016). A general framework for the annotation of causality based on framenet. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, may. European Language Resources Association (ELRA).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wolff, P., Klettke, B., Ventura, T., and Song, G. (2005). Expressing causation in english and other languages. In Woo kyoung Ahn, Robert Goldstone, Bradley C. Love, Arthur B. Markman and Phillip Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas Medin*. Washington, DC, US:American Psychological Association, pp. 29–48.

Yimam, S. M. and Gurevych, I. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In The 51th Annual Meeting of the Association for Computational Linguistics (ACL) - System Demonstrations, ACL'13, pages 1–6.