Saarland University

Department of Computational Linguistics

Master Thesis

submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computational Linguistics

---

Situation Entity Types: a Cross-linguistic Corpus Study and a Comparison of Automatic Classifiers

---

Author:
Kleio-Isidora Mavridou
Matriculation: 2529615

Supervisors:
Prof. Dr. Manfred Pinkal
Annemarie Friedrich

15 February 2016

Abstract

Smith (2003) in her theory suggests that *discourse modes*, which are different types of text passages, have distinct linguistic characteristic features. One of these is the *situation entity type* predominant in each mode. Situation entities are the *states*, *events*, *general statives* and *abstract entities* that are introduced into the discourse by clauses. Extending the work of Friedrich and Palmer (2014b), we adjust an existing English annotation scheme for use on German data and create the first parallel English-German corpus annotated with situation entities. Our aim is to find out whether situation entity types correlate cross-linguistically. To this end we conduct a corpus-based study using parallel aligned English-German texts. Similarities and differences between aligned clauses are analyzed and it is found that situation entities mostly correspond across the two languages and most mismatches are systematic. As a next step, we create the first computational model for situation entity type classification for German. In addition, we build an English classifier based on the work of Palmer et al. (2007) for the sake of comparison. We use five different simple feature sets consisting of part-of-speech, word and lemma information as well as combinations of POS tags with words and lemmata, and we achieve an absolute accuracy gain of up to 12.6% and 14% against a majority class baseline for English and German, respectively. Finally, we explore the potential of the domain adaptation approach of Daumé III (2007) by exploiting additional annotated English data for training and provide a baseline for future work using this approach.

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, 15 February 2016 Signature

Acknowledgement

First of all, I would like to express my gratitude to my supervisors, Prof. Dr. Manfred Pinkal and Annemarie Friedrich, on whose work this thesis builds, for their invaluable advice and support. Without Annemarie's directions, explanations and feedback this work would not have been possible. I am also very grateful to Alexis Palmer for her suggestions, comments and for the great contribution to this work with her pilot study on discourse modes.

I would like to thank Melissa Peate Sørensen, who along with Annemarie Friedrich ran a large-scale web experiment on the interpretation of the German perfect and provided a means to support the introduction of a new annotation label for German.

I am thankful to the annotators of our corpus Christina Bocionek, Melissa Peate Sørensen and Fernando Ardente, without the help of which I would not have been able to write this thesis.

Many thanks to Wladimir Sidorenko for his invaluable support with installing and running the discourse segmenter for German, which he provided to us.

My gratitude goes to Ilya Kornev for his contribution in solving technical problems, installing tools and running scripts whenever needed as well as for providing moral support without complaining. His dedication has saved me from despair several times.

Many thanks go to Madhumita, Nicos Bambounis, Zhe Wang and Omid Moradiannasab for the brainstorming, information exchange and for the invaluable moments we spent together.

I would also like to thank my parents and sisters for their support in every possible way throughout my Master studies.

Contents

1 Introduction

Humans understand texts by interpreting clauses, sentences and text passages in context. For a computer, this is a non-trivial task and requires knowledge of the relations between the text fragments. Discourse analysis is the analysis of language beyond the sentence boundaries. This kind of analysis is very important in natural language processing (NLP), since it has been proven to be beneficial for applications like sentiment analysis (Heerschop et al., 2011), natural language generation (Prasad et al., 2005), summarization (Louis et al., 2010) and machine translation (Meyer, 2014). Research about discourse in these applications has started only recently and though it is clear that discourse information is useful, it is less clear what the most fruitful level of analysis is.

In this thesis we address the issue of discourse analysis at the local level of the clause. Different text passage types are comprised of clauses or situations that can be categorized into groups according to their linguistic characteristics. Our main questions are whether these situations are the same across parallel texts of two closely related languages, like English and German, and whether lexico-syntactic information can provide useful features for their automatic classification in both languages.

1.1 Situation Entity Theory Overview

Texts are usefully grouped into genres which are classifications of texts, not necessarily reflecting the linguistic differences of different text types. Genres are determined on external criteria relating to the author's or speaker's purpose (Biber, 1988), and each genre represents an activity with its own structure, purpose and conventions. This reasonable categorization has led many researchers to focus on genre as the appropriate level of discourse analysis. However, there is variation within texts, even of the same genre. A news text, for instance, might start with a narrative passage, moving on to a descriptive passage and also include passages with the author's own comments and arguments. All these passages have distinct linguistic characteristics and make different contributions to the text. Smith (2003) offers a linguistic theory to the study of discourse as opposed to the pragmatic approach where genre is in the focus. The key idea of her work are the **"discourse modes"** or **DMs**, which are linguistic properties of text passages.

The different modes (Narrative, Report, Description, Information and Argument/Commentary) introduce certain types of situations into the universe of dis-

course. These situations, known as **"situation entities"** or **SEs**, are expressed linguistically at the clause level. Following Smith (2003), Palmer et al. (2007) and Friedrich and Palmer (2014b) distinguish the following types of SEs: States, Events, Generalizing Sentences, Generic Sentences, Facts, Propositions, Imperatives and Questions. Each mode shows a different distribution of situation entity types. The Narrative mode, for instance, is dominated by States and Events while a higher proportion of General Statives is present in the Information mode. Since there is an interaction between these two layers of analysis, a study at the level of situation entities can be the basis for the study of discourse modes and discourse analysis in general.

1.2 Contributions

In our work we investigate the cross-linguistic correspondence of situation entity types for two closely related languages, English and German. Our first contribution is a corpus study which we conduct in order to find correlations between SE types in English and German. For this purpose we first adapt the existing English annotation scheme of Friedrich and Palmer (2014b) for use on German data and create the first parallel corpus annotated with situation entities. Although the linguistic realization of the SE types in clauses is obviously language-dependent (Smith, 1991), the important observation of this adaptation is that the SE categories are applicable to German. We then identify and analyze the differences between SE types of aligned clauses and and their correlations across the two languages.

In a second phase we create the first automatic model for classifying clauses with regard to their situation entity types in German. The task of situation entity type classification has been addressed previously, though for English only. In addition, we build a classifier for English based on the work of Palmer et al. (2007) for the sake of comparison against our German model. We use five different feature sets containing part-of-speech, lemma and word information as well as combinations of POS with lemma and words. Since we only have a small amount of German annotated data we also investigate the domain adaptation approach suggested by Daumé III (2007), where we exploit the bigger amount of available English annotated data to improve our model using these data as additional training data.

A model for labeling clauses with regard to their situation entity types will be a useful resource for identifying the types of text passages, or discourse modes, that are present in a text. This, in its turn, will provide the basis for an alternative approach

to discourse analysis. We experiment with basic features and get substantial results; we, thus, believe that there is room for improvement by adding more sophisticated, syntactic-semantic features, which is beyond the scope of this thesis.

1.3 Results

In our corpus study we find that the aligned segments of our parallel corpus have the same labels in most cases. Mismatches are thoroughly studied and our qualitative analysis shows that half of them are a result of cross-linguistic differences and translation effects like transposition, where one word class is replaced with another without changing the meaning of the message (Vinay et al., 1995). The SE type shifts fall into eight categories and the results suggest that most of them are systematic.

We show that our simple feature sets consisting of part-of-speech, word and lemma information beat the baseline, which is defined by assigning the most frequent label in the training set to each clause, by up to 12.6% and 14% absolute for English and German, respectively. The combination of POS tags and lemmata is shown to be the most useful feature set for English, while our German classifier benefits more by words only or words in combination with POS tags.

We demonstrate that the fine-grained part-of-speech categories we use as features have a significant contribution to our classification models. By running the same experiments replacing our our fine-grained part-of-speech categories with a common tagset consisting of only 12 labels we get worse results, which on the German side even fail to beat the baseline.

In the domain adaptation experiment we treat English as a different domain and use two additional English data sets to train our German model. We only experiment with POS tags as features after mapping the English tags to the German ones. Our best results beat not only the baseline but also the results of our experiments using German data only, although the impact is small. However, we establish a baseline and suggest that in future work the POS tags could be combined with additional syntactic information, like determiner type, bare plural, etc., as well as semantic information such as tense, aspect and word sense classes which might help increase the performance of the classifier.

1.4 Structure of the Thesis

In Section 2 we give an overview of the linguistic theory of discourse modes and situation entities (Smith, 2003) and present related work. Section 3 discusses German grammatical phenomena and constructions that do not occur in English, which we took into consideration when adjusting the annotation manual of Friedrich and Palmer (2014b). Section 4 presents the parallel corpus we created and used to conduct our cross-linguistic study and to develop our classification model. Section 5 is dedicated to our corpus-based study on the correspondence of situation entities across English and German. We provide a confusion matrix of the situation entity types between the two languages and discuss the results of our qualitative analysis of the mismatches. In Section 6 we present the creation of our classification model for situation entities. We explain our used feature sets and our classification approach, and analyze the results for both our German and English classification experiments. In addition, we report on our results on the domain adaptation experiment. Finally, in Section 7 we summarize and discuss our work and provide ideas for future work.

2 Background and Related Work

2.1 Background: Discourse Modes and Situation Entities

Discourse modes are linguistic properties at the level of text passage and hold for almost all genre categories. Smith (2003) distinguishes between five different modes: Narrative, Description, Report, Information and Argument/Commentary. The following passages, introduce the different types of modes and give an intuition for the distinctions between them (examples taken from Smith, 2003).

(1) *She put on her apron, took a lump of clay from the bin and weighed off enough from a small vase. The clay was wet. Frowning, she cut the lump in half with a cheese-wire to check for air-bubbles, then slammed the pieces together much harder than usual. A fleck of clay spun off and hit her forehead, just above her right eye.* (NARRATIVE)

(2) *In the passenger car every window was propped open with a stick of kindling wood. A breeze blew through, hot and then cool, fragrant of the woods and yellow flowers and of the train. The yellow butterflies flew in at any window, out at any other.* (DESCRIPTION)

(3) *Near a heavily fortified Jewish settlement in the Gaza Strip, an Israeli soldier and a Palestinian policeman were wounded as Palestinian protests for the release of 1,650 prisoners degenerated into confrontations. Israeli military officials say they are investigating the source of fire that wounded the soldier.* (REPORT)

(4) *Thanks to advanced new imaging techniques, the internal world of the mind is becoming more and more visible. Just as X-ray scans reveal our bones, the latest brain scans reveal the origin of our thoughts, moods and memories. Scientists can observe how the brain registers a joke or experiences a painful memory.* (INFORMATION)

(5) *The press has trumpeted the news that crude oil prices are three times higher than they were a year ago. But it was the $10 or $11 price of February 1999, not the one day, that really deserved the headlines.* (ARGUMENT)

Two linguistic features characterize the modes. The first is the principle of text progression that holds for each mode. While three of the modes, NARRATIVE, REPORT

and DESCRIPTION have temporal principles of progression, INFORMATION and ARGUMENT are atemporal modes and progress metaphorically through the domain of the text. The second characteristic is the type of situation that a text passage introduces to the discourse. Situation entities are conceptual categories, realized at the clause level. They fall into four broad categories motivated by patterns in their linguistic behavior. There are certain features that can help to identify the situation entity type of a clause: lexical aspect, habituality and genericity of the main referent. Based on Smith (2003), Friedrich and Palmer (2014b) distinguish the following SE types:

Eventualities. There is no agreed-upon system for the classification of lexical aspect. Ryle (1949) distinguishes between accomplishments and achievements, while Kenny (1963) talks about states, activities and performances. The most influential distinction, the one of Vendler (1957), suggests a four-way classification into states, activities, achievements and accomplishments, whereas Bach (1986) classifies eventualities into states, processes and events. Friedrich and Palmer (2014b) make the distinction between stative (=something is the case) (6) and eventive (=something happens) (7) lexical aspect only. In addition to STATES and EVENTS, they introduce the type REPORT (8) as a subtype of EVENT, which is introduced by situations containing verbs of speech.

(6) *"John is cute", (*STATE*)*

(7) *said Mary. (*REPORT*)*

(8) *John won the race. (*EVENT*)*

General Statives. This category includes GENERALIZING SENTENCES (9), which express regularities about the world (Carlson, 2005) related to a non-generic main referent, and GENERIC SENTENCES (10), which make statements about kinds or generic concepts or motions (Krifka et al., 1995).

(9) *Mary often goes to the gym. (*GENERALIZING*)*

(10) *The lion has a bushy tail. (*GENERIC*)*

Abstract Entities. This class comprises FACTS (11), which are objects of knowledge, and PROPOSITIONS (12), which are objects of belief.

(11) *Mary knows that John went to the movies. (*FACT*)*

(12) *Mary believes that John lied to her. (*PROPOSITION*)*

Speech Acts is the last class of SE types and includes clauses containing QUESTIONS (13) and IMPERATIVES (14).

(13) *How was the movie? (*QUESTION*)*

(14) *Give me your phone! (*IMPERATIVE*)*

Smith (2003) predicted that different types of situation entities predominate in different modes. The predicted predominant situation entity types per discourse mode are summarized below:

Narrative - eventualities
Report - eventualities and general statives
Description - states, ongoing events
Information - general statives
Argument - abstract entities, general statives

In our cross-linguistic corpus-based study we will explore whether the predictions apply to our data, both English and German. Before this, we want to test the suitability of the situation entity types for use on German data. Finally, we want to create automatic models to automatically label English and German clauses with regard to their SE type. Such models already exist for English but to our best knowledge, this is the first attempt to do this for German.

2.2 Empirical Studies on Situation Entities

Xue and Zhang (2014) use an annotation approach similar to that of Friedrich and Palmer (2014b), focusing only on eventualities. They use a "distant annotation" method, with which they mark tense and modality on the English part of a Chinese-English word-aligned parallel corpus and then project the annotations onto the Chinese part. Each text span is annotated with regard to semantic tense, modality and event type. The event types they distinguish are habitual event, state, on-going event, completed event, and episodic event.

Influenced by the work of Smith (2003), Palmer and Friedrich (2014) investigate the relationship between situation entities and various text genres, with the aim to compare the predictions of Smith (2003) with evidence from text corpora. Due to the lack of data labeled with discourse modes, they use different genre categories as proxy for discourse modes, for example news text as a proxy for the REPORT mode. The theory

is tested on two corpora containing documents of different genres, specifically news, letters, essays and jokes. The analysis shows that the different genres have indeed different SE type distributions, which correspond to the predictions of Smith (2003).

Extending the work of Palmer and Friedrich (2014) and with the purpose of testing the theory of Smith (2003), we conduct a cross-linguistic study on discourse modes and situation entities in a preliminary version of this work. We conduct two separate studies on discourse modes and situation entities to find out the cross-linguistic correspondence of DMs and SE types for the language pair English-German. To this end, we create the first corpus labeled with DMs and the first corpus annotated with SE information and adapt the existing English annotation scheme of Friedrich and Palmer (2014b) to German. The two studies are then combined to test the predictions of Smith (2003). The results mostly confirm the predicted distributions of SE types per DM in both languages and show that DMs and SE types are most of the times the same across the two languages. Differences are analyzed and it is found that most of them result from translation effects (Mavridou et al., 2015). In this thesis we extend the analysis by adding more data that support our preliminary results and reveal additional insights.

2.3 Computational Models for the Study of Situation Entities

Since discourse modes and situation entities are interacting levels of linguistic analysis, the creation of automatic methods for labeling clauses with their situation entity types would be useful for automatically identifying the types of text passages. Though situation entities are well-studied in linguistics, computational models for their study have emerged only recently.

Some previous work has addressed the classification of clauses according to their SE type. Palmer et al. (2004) assign clauses with phrase structure trees and f-structure representations and augment these with lexicosemantic information. The enriched parses are then passed to a transfer system, used for applying linguistic tests. These tests are ranked according to their strength as correlates of particular situation entities. Results show that inclusion of lexical information improves recall while at the same time decreasing precision.

Palmer et al. (2007) are the first to create a classification model for situation entities. They use various types of linguistically-motivated and deep syntactic features in two probabilistic models for SE type classification: a labeling model based on clause-level features as well as a sequencing model that takes into account discourse patterns

between the clauses (up to six previous labels). The results indicate that the incorporation of local context increases performance though the best results (no more than two labels) show an accuracy of round 54%.

Zarcone and Lenci (2008) create two computational models to automatically identify event types in Italian, using adverbial, morphological and syntactic features. In the first model, the identification of event types is modelled as a supervised classification task, performed with a Maximum Entropy classifier. Occurrences of 28 Italian verbs are classified according to the the Vendlerian verb classes (states, activities, accomplishments and achievements). The model outperforms the baseline and shows that contextual features help to identify event types. The second model uses Self-Organizing Maps to identify event types in an unsupervised manner. 40 verbs are represented as distributional vectors recording their co-occurence frequencies with contextual features and an accuracy of 72.5% is achieved.

In Section 2.1 we have mentioned the annotation scheme of Friedrich and Palmer (2014b), where clauses are categorized into four broader categories: eventualities, general statives, abstract entities and speech acts. Similarly, Xu and Huang (2014) provide a Chinese corpus, in which sentences are labeled as speech acts, events or modalities, the latter corresponding to the coercion triggered by modality as described in our annotation scheme. Their notion of events includes the four Vendler classes (Vendler, 1957), the type semelfactive introduced by Smith (1991) as well as finer-grained event categories. Two classification experiments using a SVM are conducted, one involving only the three main event categories and the other incorporating the finer-grained types. The results of the coarse level classification show an accuracy of 83.6% using, while the best result on the finer-grained level is 62.1%.

Using a corpus annotated based on the scheme of Xue and Zhang (2014), Zhang and Xue (2014) experiment with three approaches on the automatic inference of Chinese semantic tense. In the first experiment, automatically derived modality and eventuality type are used as features in tense inference. The second experiment involves joint learning on tense and each of these features, and the last uses neural networks to train models for tense prediction. All three approaches outperform the baseline, although accuracy is higher on newswire text.

Other related works handle tasks related to the features we annotate. Siegel and McKeown (2000) address the classification of the aspectual category of clauses, i.e. whether a verb is used in a stative or dynamic sense. A verb's aspectual class can be predicted with the use of certain linguistic indicators, which are co-occurence frequencies between the verb of a clause and certain linguistic phenomena, such as the

progressive or the perfect tense. Siegel and McKeown (2000) use three supervised and one unsupervised machine learning approach for automatic aspectual classification. For the supervised classification, around 98,000 manually parsed clauses from medical discharge summaries are used to extract frequencies for verbs according to 14 linguistic indicators. Logistic regression, decision trees and genetic programming are compared regarding to their ability to combine the linguistic indicators and the best approach achieves an accuracy of 93.9% versus a baseline of 83.8%. The unsupervised clustering algorithm is tested on a small set of 56 frequent verbs and is able to distinguish stative and event verbs.

Friedrich and Palmer (2014a), following Siegel and McKeown (2000), go one step further and treat the problem as a three-way classification task predicting the aspectual class of verbs in context as dynamic, stative or both. They introduce two data sets containing clauses labeled with aspectual category and explore linguistic indicators, distributional and instance-based features in a semi-supervised setting. For seen verbs, no feature combination achieves significant improvements over a baseline of memorizing the most frequent class of verbs. In all experimental settings, improvements are achieved by integrating instance-based features, though the other two feature sets successfully predict the aspectual class of unseen verbs.

Similarly, Friedrich and Pinkal (2015a) provide an automatic approach of classifying the aspectual category of clauses as being static, episodic or habitual. They use context-based features, partially proposed by Mathew and Katz (2009), as well as the type-based linguistic indicators suggested by Siegel and McKeown (2000). A joint model that runs a three-way classification task, but also a cascaded model that distinguishes between static and non-static as well as between episodic and habitual, are applied. The results show that the importance of the different features differs according to the subtask and that the cascaded approach is more robust for the three-way classification. Overall, high accuracies up to 80% for the three-way classification task and up to 85% for the subtasks can be reached, even for unseen verbs.

Other work focuses on the identification of generic expressions in texts. Reiter and Frank (2010) use a wide range of NP-level and sentence-level features as well as syntactic and semantic features to train a supervised classifier for identifying generic noun phrases. They use the ACE-2 corpus (Mitchell et al., 2003) and experiment with a balanced and an unbalanced data set to avoid bias effects. Their results show that all features contribute important information, with syntactic features being the most informative.

Friedrich and Pinkal (2015b) present a novel state-of-the-art approach for automati-

cally identifying generic expressions. Based on and comparing against the work of Reiter and Frank (2010) who train a supervised classifier to identify generic noun phrases, sequences not only of noun phrases but also of clauses are labeled as being generic or non-generic. For this purpose, they create the WikiGenerics corpus (Friedrich et al., 2015), balanced to contain many generic and non-generic clauses from various domains. For the classification of the instances they use information about each instance as well as genericity information of neighboring instances. The model outperforms previous approaches and the results show that the integration of context increases accuracy.

Generally, a lot of work is being done on the broader field with most of the work focusing on the automatic processing of events, like for example event identification (Saurí et al. (2005), Bethard and Martin (2006), Hongye et al. (2008)) or event extraction (UzZaman and Allen (2010), Marovic et al. (2012)).

3 Adaptation of Annotation Scheme to German

An important first step of our work is the creation of an annotation manual for German. To this end we adapt the existing English annotation scheme of Friedrich and Palmer (2014b) for use on German data. At a first stage, several paid annotators, German native speakers and students of computational linguistics, are asked to annotate German documents using the guidelines of the English scheme and report on problems and difficulties. The comments are analyzed and the scheme is then adapted to German to account for all identified cases that differ from English.

The guidelines regarding the annotation of the situation-related features (genericity of main referent, aspectual class, and habituality) apply for the German annotation scheme, as they do for the English scheme of Friedrich and Palmer (2014b). In the following sections we present cases that appear in German and need special handling because they do not occur in English. We describe how to mark the genericity of the main referent in German, especially in constructions in which the main referent is not the grammatical subject of the clause, and we talk about the German-specific derived situation entity types. Moreover, we introduce a new subtype of EVENT, namely the EVENT-PERFECT-STATE to mark clauses containing perfect tense, that cannot be described neither as STATES nor as EVENTS.

3.1 Genericity of Main Referent

In order to find the main referent of a clause, the question we shall have in mind is what the sentence is about. Usually the main referent is realized as the grammatical subject of the clause. However, this is not true for all clauses and such cases are frequent in German.

One such example is the impersonal passive, which can be formed in German (in contrast to English) for intransitive verbs. During the formation of the passive clause, the pronoun "*es*" serves as a placeholder for the subject at the beginning of the clause and can be omitted if another element of the clause takes up this position (Hentschel, 2003). Since there is no grammatical subject, annotators have to decide whether the implied main referent of the clause is a specific person/group/organization, as in (15) and (17), or whether the sentence is about a kind or an abstract individual/entity, as in (16), where the implied main referent are all people who sew in general.

(15) *Jetzt ist Pause, (non-generic,* STATE*)*
es wird wieder geredet. (non-generic, EVENT*)*

(16) *Früher gab es keine Nähmaschinen, (generic,* Generic Sentence*)*
heute wird anders genäht. (generic, Generic Sentence*)*

(17) *Abends wird im Gemeinschaftsraum gespielt, getrunken und gefeiert. (non-generic,* Generalizing Sentence*)*

"*Es*" is often used as an expletive. Expletives often serve as a way to rhematize the subject and are themselves semantically empty (Hentschel, 2003). In examples (18) and (19), the pronoun functions as a placeholder and can be omitted, the main referents are "*ein Bild*" and "*der Flieder*" respectively.

(18) *Es hängt ein Bild an der Wand. (non-generic)*

(19) *Es blühte der Flieder. (non-generic)*

There is a group of impersonal verbs that express the perceptions of a person, which are usually expressed with stative verbs, and require an argument in dative, as in (20), or accusative, as in (21). In both cases, "es" is only used if the first position of the clause is empty (Hentschel, 2003). In these cases, the argument in dative or accusative is considered to be the main referent of the clause.

(20) *Es gruselt mir vor dir. (non-generic)*

(21) *Mich friert es. (non-generic)*

In other cases, "*es*" is obligatory and is considered by many as the subject of the clause (Helbig and Buscha, 2001). This is the case with constructions like *gehen um*, *sich handeln um*, *kommen zu*, etc. Here, even if "*es*" is the formal subject, we consider the main referent to be the person, situation or object that the clause is referring to. This is also true when "*es*" appears with verbs that express existence. These clauses are analogous to the existential clauses in English.

(22) *Es handelt sich um ein einen Notfall. (non-generic)*

(23) *Derzeit gibt es viel zu tun. (non-generic)*

(24) *Es war einmal ein Drache. (non-generic)*

Annotators are asked to annotate the pronoun "*es*" as **no-main-referent** in clauses where no other main referent can be identified, for example with verbs expressing natural phenomena and especially those describing the weather. In these cases, "*es*" is obligatory and cannot be replaced with anything else or omitted (Hentschel, 2003).

(25) *Es grünt und blüht in Wald und Flur. (**no-main-referent**,* State*)*

(26) *Letzte Nacht regnete es stark. (**no-main-referent**,* Event*)*

(27) *Es ist kalt. (**no-main-referent**,* State*)*

Another interesting case for German is the indefinite pronoun "*man*". It is analogous to the English *people*, *they*, *one*, etc., and refers either to a specific group of people, like in example (28), or to abstract individuals or entities, like in example (29). In these cases annotators shall distinguish whether the clause is about a generic or non-generic entity.

(28) *Jetzt ist Pause, man redet wieder. (non-generic,* Event*)*

(29) *Heute näht man anders. (generic,* Generic Sentence*)*

3.2 Derived Situation Entity Types

In some cases, the situation entity type of a clause changes due to the presence of some linguistic indication of uncertainty or doubt about the status of the situation, or by the presence of linguistic information that focuses on the post-state of an event. These cases are referred to as *derived situation entity types*.

The linguistic features which trigger this change in the type of situation entity present in a clause are negation, modality, future tense, conditionality as well as subjectivity. When these features occur, they cause a coercion of a clause that would otherwise be marked as Event to the category of State. Coercions only occur for the Event type and for no other type.

In this Section, we discuss features that trigger a coercion, which do not occur in English. Our main focus, however, is the perfect, which can have both a stative or an event reading, or even be underspecified depending on the context. For this purpose we introduce Event-Perfect-State, a new type to mark clauses that can not be labeled neither as States nor as Events.

3.2.1 Perfect

In English, we consider clauses in present perfect and past perfect to be states (Katz, 2003), as they focus on the circumstances of an action being completed at the time of

reference. In German, however, this rule does not apply since it is not strictly defined when to use the present perfect (*Perfekt*) or the simple past (*Präteritum*). These two tenses can often be used interchangeably (especially in spoken language) to describe events that were completed in the past. Examples (30), (31) and (32) show the use of the present perfect for describing past events.

(30) *Schiller hat "die Räuber" im Jahre 1781 geschrieben. (*EVENT*)*

(31) *Gestern sind wir alle ins Kino gegangen. (*EVENT*)*

(32) *Ich möchte allen danken, (*STATE*)*
*die so intensiv an diesem Thema mitgewirkt haben. (*EVENT*)*

Like in English, however, clauses in present perfect and past perfect can also be considered to be states in German, as they can also focus on the outcome or the state reached after a process and have an aspectual reading (Klein, 2000). In example (33), the interpretation of the clause is that there is now snow around, in example (34) that the speaker is not hungry and in example (35) that the person was aswoon when she was found.

(33) *Guck mal, es hat die ganze Nacht geschneit. (*STATE*)*
*(In contrast: Es schneite die ganze Nacht. (*EVENT*))*

(34) *Ich habe schon gegessen. (*STATE*)*

(35) *Sie hatte das Bewusstsein verloren, (*STATE*)*
*als ihre Retter sie endlich befreiten. (*EVENT*)*

In contrast to English, it is not clear in German where the boundary lies between STATE and EVENT in clauses containing perfect . Sentences differ in the degree to which they invoke an eventive or a stative interpretation: some sentences seem to focus more strongly on the after-state of the event than others. In many cases, it is a matter of underspecification that makes it difficult to decide whether the clauses express STATE or EVENT, the sentence simply has both functions. For this reason, we have introduced a new subtype of EVENT, the EVENT-PERFECT-STATE. When the clause is underspecified and annotators feel that a clause is neither primarily a STATE nor primarily an EVENT, they are asked to mark them with the EVENT-PERFECT-STATE subtype.

(36) *Das scheint ein wichtiges Thema zu sein, und ich sage Ihnen das als jemand,*
der sich selber damit ***beschäftigt hat****. (*EVENT-PERFECT-STATE*)*

(37) *Ich habe die Stelle! Sie* ***haben*** *mir den Job* ***gegeben****. (*Event-Perfect-State*)*

(38) *Bitte, setzen Sie sich, ich* ***habe*** *schon zwei Bier* ***bestellt****. (*Event-Perfect-State*)*

To support our decision to introduce this new label we conduct a large-scale web-based experiment where we ask annotators, native speakers of German, to rate in a scale from 1 to 5 whether the state or the event matters more for a target word in a sentence. We use reference clauses that we have labeled beforehand as State, Event and Event-Perfect-State, all appearing in the same proportion in the texts we present to the annotators. Our hypothesis is supported by the results since most reference labels we had assigned to clauses match the ratings of the annotators who indeed recognize a third reading of the perfect aside from the stative and the event reading (Mavridou et al., 2015).

3.2.2 Subjunctive

There are two different types to express the subjunctive mood (*Konjunktiv*) in German: The Subjunctive I (based on the present tense) and the Subjunctive II (based on the past tense). Contrary to the indicative, the subjunctive is used to express doubt, possibility, speculations, conditionality and other "unreal" conditions (Bihl, 1949). When the subjunctive is used, situations containing verbs with dynamic aspectual class are marked as States, although this coercion does not apply to General Statives, as in examples (41) and (42).

(39) ***Hätten*** *wir das Geld, (*State*)*
gingen *wir morgen schon im Urlaub. (****dynamic*** *aspectual class,* ***static****, conditionality →* State*)*

(40) *Ein plötzliches Zerreissen eines Tisches und sauberes Zerspringen eines Brotmessers* ***habe*** *er* ***beobachtet****. (****dynamic*** *aspectual class,* ***static****, modality ≈ doubt →* State*)*

(41) *Sie sagte, (*Report*)*
sie ***tue*** *das regelmäßig um Menschen in Not zu helfen. (****dynamic*** *aspectual class,* ***habitual****,* Generalizing Sentence*)*

(42) *Er glaubte, (*State*)*
der Mensch ***arbeite*** *nur aus Zwang. (****dynamic*** *aspectual class,* ***habitual****,* Generic Sentence*)*

3.2.3 Statal Passive Voice

The statal passive focuses, as its name denotes, on the result or the "state" reached after a process (Bihl, 1949). Clauses in statal passive are therefore marked as States, as in examples (43), (44) and (45).

(43) *Die Tür ist geöffnet. (*State*)*

(44) *Der Brief ist geschrieben. (*State*)*
*(In contrast: Der Brief ist geschrieben worden. (*Event*)*

(45) *Sein Herz war für Melanie Nowara entflammt. (*State*)*

3.2.4 Final Clauses with "damit"

Final clauses are used to describe a purpose, an intention or a goal. Clauses starting with "damit" that contain events are coerced to States, as they are considered to describe the possibility of an event, like in examples (46) and (47). Again, this coercion does not apply to General Statives, as we see in (48).

(46) *Es wird Druck ausgeübt, (*Event*)*
*damit die Pharmaindustrie neue Impfstoffe entwickelt. (**dynamic** aspectual class, **static** →* State*)*

(47) *Erinnere mich nochmal, (*Imperative*)*
*damit ich pünktlich komme. (**dynamic** aspectual class, **static** →* State*)*

(48) *Damit sie pausenlos arbeiten, (*Generalizing Sentence*, **habitual**)*
*werden sie ständig überwacht. (*Generalizing Sentence*, **habitual**)*

3.2.5 Modal Constructions

Modality can trigger a change in the type of the situation entity. It is not only expressed through the modal verbs (in German: *müssen, sollen, dürfen, können, mögen, wollen*) but also through various grammatical constructions. When these constructions appear in clauses that would be normally marked as Events, the situation entity type is coerced to State. Two common structures are described in this section, both of them used as alternatives to the passive voice.

haben/ sein + zu + infinitive. Phrases containing this construction are used to denote necessity or possibility and, thus, are marked as STATES.

(49) *Sie **haben** es **zu unterlassen**, vertrauliche Informationen weiterzugeben. (necessity → STATE)*

(50) *Sie **hat** etwas **zu verschenken**. (possibility → STATE)*

(51) *Diesem Kommentar **ist** nichts **hinzuzufügen**. (possibility → STATE)*

(52) *Die Wohnung **ist** beim Auszug **zu renovieren**. (necessity → STATE)*

sich lassen + infinitive. This construction can be analyzed with the semantics of the verb *can*. Here again, situations containing verbs with dynamic aspectual class are coerced to STATES.

(53) *Den Beteiligten muss klar sein, (STATE)*
*dass **sich** dieser Konflikt ohne Kompromisse **lösen lässt**. (**dynamic** aspectual class, **static**, possibility → STATE)*

(54) *Der Umweltskandal **ließ sich verbergen**. (**dynamic** aspectual class, **static**, ability → STATE)*

A coercion does not apply in the case of general statives.

(55) *Mit dieser Software **lassen sich** täglich Millionen von Tweets **auswerten**. (GENERALIZING SENTENCE)*

(56) *Geschickt **lassen sich** jetzt aus Gegnern Freunde machen. (GENERIC SENTENCE)*

Note that it should not be confused with another *sich-lassen* construction, which requires an animated subject and has the meaning of "allow", as in (57) or "prompt", as in (58).

(57) *Anna ließ sich nicht unter Druck setzen. (= Sie ließ nicht zu, dass ...). (EVENT)*

(58) *Er ließ sich die Haare schneiden. (Er veranlasste, dass ...). (EVENT)*

4 Corpus data

In this work we explore the cross-linguistic relation of situation entity types and build two classifiers for two different languages, namely English and German. It is, therefore, necessary to have a corpus of parallel annotated data. The corpus we create consists of texts of various genres like news articles, political speeches, economy texts, TED talks and novels, all chosen to cover various topics and have different distributions of situation entity types. The corpus is made up of two main document sets to which we refer as "set 1" and "set 2". Set 1 is used for our preliminary corpus study (Mavridou et al., 2015), while the documents in set 2 were later collected for the needs of our classification task that requires more data.

4.1 Cross-linguistic corpus study

For the first part of our work parallel aligned data from different genres are required, in order to show whether situation entity types correspond cross-linguistically. For this purpose, 11 parallel English-German texts from various sources were collected: sections from the first chapters of the novels "Alice's Adventures in Wonderland" and "Anna Karenina" from the OPUS "Books" collection (Tiedemann, 2012), three documents from a customized version of the Europarl corpus (Koehn, 2005) created for translation studies (Islam and Mehler, 2012), two texts from the News Commentary corpus (part of the WMT 2013 shared task training data[1]) and two articles from Global Voices Online, a multilingual news website[2]. In addition, two documents ("Sophie's World" and "Economy Texts") from the Smultron corpus (Volk et al., 2015) were used. The source language of the documents is either English, German or a third language into which these texts were translated, like Norwegian and Russian. Table 1 shows the number of English, German and aligned clauses we use to conduct our preliminary cross-linguistic study described in Mavridou et al. (2015). In this thesis we use our entire document set described below.

[1] http://statmt.org/wmt13/

[2] https://globalvoices.org/

	# English clauses	# German clauses	# aligned clauses
Europarl	355	322	295
News texts	355	340	303
Novels	189	184	163
Smultron	1794	1677	1028
Total	**2693**	**2523**	**1789**

Table 1: Number of English, German and aligned clauses per corpus section in set 1.

4.2 Classification of situation entity types

For the classification task of SE types more annotated data and from various genres are required to avoid same patterns and distributions of situation entity types. 19 further documents were collected: seven news texts from the Project Syndicate website[3], excerpts from the novels "Jane Eyre", "Madame Bovary" and "The Metamorphosis" from the OPUS corpus, with English, French and German as source languages, respectively, three short stories and seven TED talks[4] covering various topics, like language evolution, breast cancer detection, and sea pollution. The number of English, German and aligned clauses in set 2 are summarized in Table 2.

	# English clauses	# German clauses	# aligned clauses
News texts	479	479	394
Novels	1115	1055	816
Short stories	349	339	289
TED talks	2040	2017	1707
Total	**3983**	**3899**	**3206**

Table 2: Number of English, German and aligned clauses per corpus section in set 2.

[3]http://www.project-syndicate.org/
[4]https://wit3.fbk.eu/

4.3 Overall corpus

The first results of our empirical corpus study are presented in Mavridou et al. (2015). Here we extend this work by using the total amount of aligned clauses to explore whether the additional data support our first observations. Moreover, we use all 30 documents to build our English and German classifiers. The total amount of English, German and aligned clauses per section in our corpus is shown in Table 3.

	# English clauses	# German clauses	# aligned clauses
Europarl	355	322	295
News texts	834	819	697
Novels	1304	1239	779
Short stories	349	339	289
Smultron	1794	1677	1028
TED talks	2040	2017	1707
Total	**6676**	**6422**	**4995**

Table 3: Total amount of English, German and aligned clauses per corpus section.

Our data were chosen to not only cover different topics and genres but also to have a fair proportion of generic sentences. However, we cannot say that the amount of eventualities and general statives in our corpus is balanced; eventualities, especially states, dominate in our documents. The distributions of situation entity types in our English and German data are shown in Figure 1 and Figure 2, respectively.

We see that both languages have similar distributions, although STATES are more prevalent in the English part, mainly due to the perfect coercion. Another interesting finding is that GENERIC SENTENCES are almost twice as many as in the German part, which suggests that German primes the generic reading of clauses. Indeed, as annotators of both languages, we ourselves could observe this tendency in our own annotations.

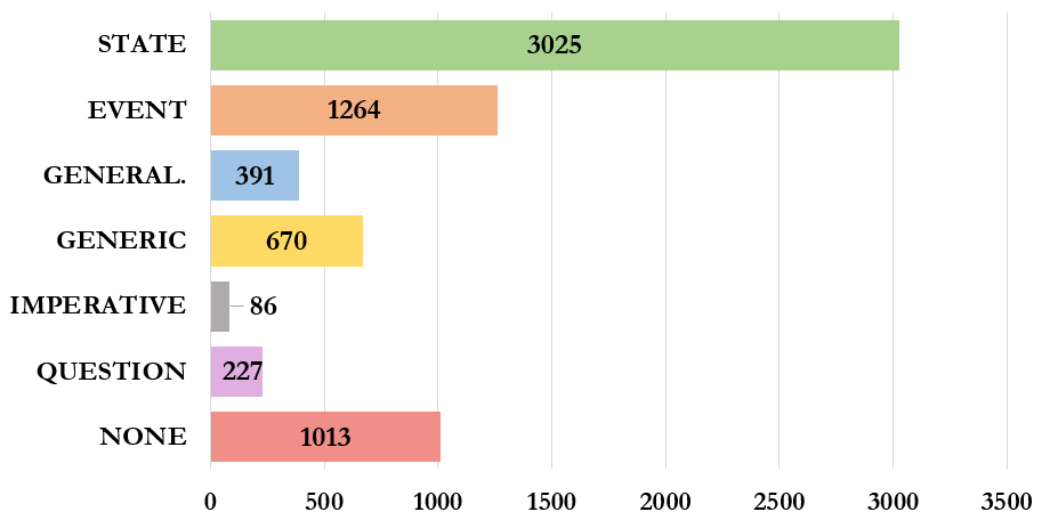

Figure 1: Distribution of situation entities in the English data.

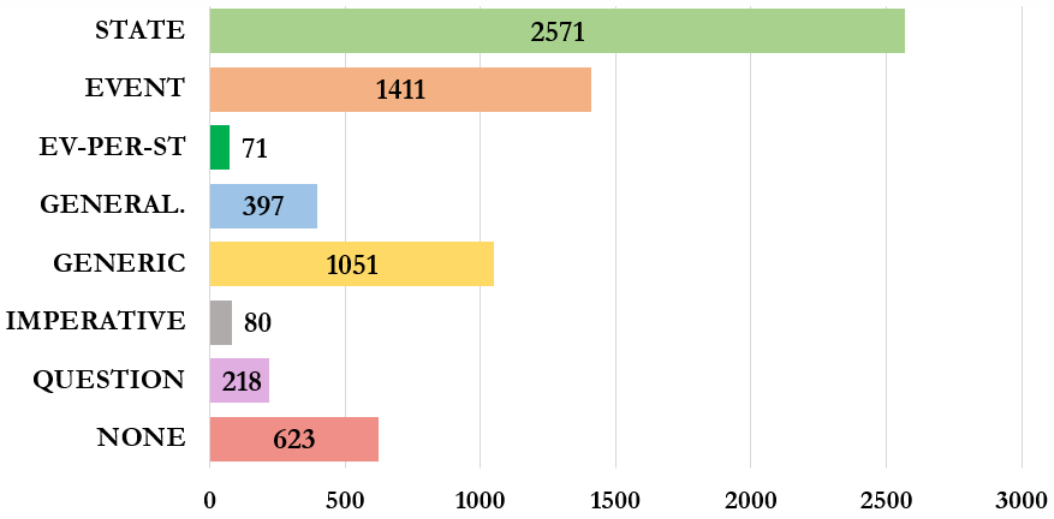

Figure 2: Distribution of situation entities in the German data.

4.4 Segmentation

Situation are introduced to the discourse by clauses and for this reason we use clauses as our basic units for annotation. Since sentences are usually made up of multiple

clauses, we need a way to segment the sentences into the clauses they are comprised of.

The English part of our data is segmented using SPADE (Soricut and Marcu, 2003), a freely available discourse parser. The integrated discourse segmenter takes a one-sentence-per-line text as input, parses the text internally and outputs the clause-segmented text. Since the output of the segmenter is sometimes more fine-grained than required, some post-processing steps are applied to avoid verbless segments. The sentences are either extracted directly as such from the XML files of our corpus or the texts are split after periods and are manually post-processed wherever necessary (mainly in the case of abbreviations or ellipses).

For the German part we use a syntax-based discourse segmenter[5], which requires dependency-parsed text as input. The dependency parses are manually created using BitPar (Schmid, 2004). However, paragraph marks are not properly handled, at least by early versions of the tool. To overcome this issue, paragraphs in the German text are replaced by dashes before tagging and segmenting.

4.5 Annotation and Agreement

Each segment is annotated by three different paid annotators in each language, all students of computational linguistics, including the author, who annotated the text in both languages. Annotators receive the existing extensive English manual of Friedrich and Palmer (2014b) plus the adapted German manual, and are trained on documents not included in the corpus. The Smultron part of the data is annotated by a different combination of annotators than the rest of the corpus. The gold standard labels are created via majority voting and contain the cases where at least two of the annotators agree on the certain label.

The agreement between the three annotators is calculated using Fleiss' κ (Fleiss' kappa). Fleiss' κ is a statistical measure for calculating the degree of agreement between more than two annotators and takes into account the amount of agreement that can be expected by chance.

Substantial agreement is achieved for both languages on document set 1, which is used to conduct our preliminary cross-linguistic study (Mavridou et al., 2015). The inter-annotator agreement for this part of our corpus is shown in Table 4.

[5]https://github.com/WladimirSidorenko/DiscourseSegmenter

	English	German
set 1 - Smultron	0.63	0.62
set 1 - Other	0.61	0.67

Table 4: Inter-annotator agreement on set 1 (Fleiss' κ).

The results of the inter-annotator agreement between the annotators for set 2, which includes news texts, short stories, TED talks and novels, is shown in Table 5. We measure agreement for this document set using Cohen's κ. This measure, as opposed to Fleiss' κ, is used to calculate the agreement between two annotators only. In this part of our corpus two annotators annotate all documents and a third annotator only labels the segments where the two main annotators disagree.

	English	German
set 2	0.56	0.68

Table 5: Inter-annotator agreement on set 2 (Cohen's κ).

The inter-annotator agreement for the English part of this section is only moderate, while for all other corpus sections we get substantial agreement. It is worth mentioning that the documents in set 2 are more challenging to annotate as a big part are literary texts which contain more difficult language and structures than other text types. We should note that two of the annotators are fluent speakers of English but not native speakers which might also influence the results. Moreover, all TED talks contain a high proportion of generic and generalizing sentences which are more difficult to annotate and often cause confusion. In Table 6 we see that annotators agree on most cases except in the labeling generics. Interestingly, annotator A labels more clauses as State, whereas B uses more Generic Sentences. We thus observe that the cases of disagreements between Generic Sentence and State as well as Generalizing Sentence are more than the cases of agreement on the label. This disagreement affects the overall agreement between the two annotators and explains the low κ score.

		annotator A						
		State	Event	General.	Generic	Imp.	Quest.	-
annotator B	State	1492	62	19	22	0	3	16
	Event	131	629	24	2	0	2	2
	General.	91	71	174	13	0	0	7
	Generic	295	50	118	58	0	0	5
	Imp.	5	3	0	0	54	2	3
	Quest.	40	14	2	0	0	87	1
	-	106	57	16	5	4	9	288

Table 6: Confusion matrix between the two annotators for the English side of set 2.

4.6 Alignment

After annotation, the English and German segments are aligned. The documents in set 1 are all manually aligned, except for the Smultron documents. During the manual alignment we take into account segments that generally match, even if they differ in their construction, as long as the main verb is the same or they refer to the same main referent and convey the same meaning. For example, the following clauses are aligned because, despite of the different lexical choice, they have the same subject and make an equivalent statement:

(59) *She was startled. (English)*
Sie fuhr zusammen. (German)

Moreover, a situation entity label in at least one of the two languages must be present, in order for the segments to be aligned. Segments that contain the same information and even the same verbs but differ in their main verb, are ignored during the alignment. This is illustrated with the following pair, where the main verb "is" in the English sentence does not appear in the German sentence:

(60) *What I want to say* ***is*** *that it is possible. (English)* — *Ich will sagen, dass es möglich ist. (German)*

For the Smultron part, on the other hand, we conduct a semi-automatic alignment. The Smultron documents come with token and sentence-level alignments but these

do not always overlap with the clause segmentation produced by our segmenters. For the purpose of alignment, the main verb of each English segment is identified with the help of dependency parses (Klein and Manning, 2003). This English segment is then aligned to the German segment that contains the verb to which the English verb is aligned according to the given information. Again, segments whose main verbs don't correspond are ignored.

However, as Table 1 shows, a high number of clauses is not aligned during the alignment of the Smultron corpus. One reason for this, the one that causes most of the losses, is that verbs which convey the same meaning but are not direct lexical equivalents are in most of the cases not aligned in the corpus. So, for example, there is no verb alignment between the clauses:

(61) *...further reduced profit in the quarter. (English)*
...wirkten sich ebenfalls negativ auf den Quartalgewinn aus. (German)

Furthermore, since one part of the corpus are economy texts with a specific structure, there are many impersonal constructions without a verb in either of the languages, most commonly in German. An example for this is following:

(62) *Although we took sizable provisions to improve the longer-term profitability of our transformer business... (English)*
Trotz hoher Rückstellungen für Rechtskosten und für die Verbesserung der langfristigen Rentabilität unseres Transformatorengeschäfts... (German)

We use a simple alignment approach that certainly allows for improvements. First of all, due to cross-linguistic differences and translation effects, verbs in parallel texts do not always match in form. Translators, though, often use not only direct equivalents and synonyms but also paraphrases and different structures that carry the same meaning. For the alignment of such structures it should be possible to go beyond word-based alignments and allow for phrase alignments by exploiting resources to derive semantic relatedness, which is beyond the scope of this thesis.

5 Empirical Cross-linguistic Corpus Study

We aim to answer two questions in the first part of our work: do parallel, aligned clauses contain the same situation entity types? And in which cases do the situation entity types differ cross-linguistically? After the adaptation of the English scheme to German we conduct a cross-linguistic corpus study to investigate whether SE types correlate across two closely related languages like English and German. Mismatches between the parallel clauses are analyzed and most shifts are found to be systematic.

5.1 Analysis and Results

5.1.1 Cross-linguistic Correspondence of Situation Entity Types

As a first step we create a confusion matrix to visualize the cross-linguistic correspondence of the SE types in our aligned clauses. Table 7 shows the results; the aligned segments have the same labels in most of the cases. There are some major confusions, however, which are all thoroughly analyzed.

We can see that State is the most dominant type and the label that is mostly confused, mainly with Eventand Generic Sentence. The confusion matrix also shows that German has a higher tendency to Generic Sentences, which in many cases are labeled as States in English. Imperatives and Questions, on the other hand, correspond in most of the cases across the two languages.

		German							
		State	Event	Ev-Per-St	General.	Generic	Imp.	Quest.	-
English	State	1782	230	47	48	163	3	6	105
	Event	97	910	19	30	21	1	0	27
	General.	20	25	0	194	68	1	0	18
	Generic	61	3	1	7	486	0	0	11
	Imp.	4	4	0	0	0	57	0	4
	Quest.	10	0	0	1	1	3	157	7
	-	103	50	2	33	105	1	7	60

Table 7: Confusion matrix of SE type labels between parallel English and German texts.

5.1.2 Qualitative Analysis of Mismatches

For the analysis of the differences between the two languages we extract all aligned clauses that have a different SE label in English and German. We found that the proportion of language-pair dependent and language-pair-independent mismatches between SE types does not differ much (54% and 46%, respectively) and that these differences fall into eight categories. In Table 8 we summarize all mismatch types that we identified during the analysis.

mismatch type	**# of mismatches**
language-pair dependent	671
involving perfect	172
lexical choice	216
grammatical structure	21
segmentation	213
language-pair independent	724
genericity of main referent	298
habituality	49
lexical aspectual class	60
errors and disagreements	317
Total	**1346**

Table 8: Types of mismatches.

To start with the language-independent shifts, judging the genericity of the main referent of clauses has been found to be a difficult decision (Friedrich et al., 2015). The interesting finding is that there are cases where a certain form primes a particular reading. According to the numbers in more than 75% of all cases it is the German part that primes the generic reading. (63) serves as an example for this observation.

(63) *Terrorists may also benefit.* (GENERIC SENTENCE)
Auch die Terroristen könnten profitieren. (STATE)

Another important class of language-independent differences is the ambiguous aspectual class of some verbs like *continue*, *show*, *think*, *support* or *wonder* in both languages. This kind of verbs may have a different aspectual class in different contexts,

and this is something that often confuses annotators. The following example pair illustrates this, where the English verb *show* seems to prime the dynamic reading in contrast to the German verb *zeigen*:

(64) *... they showed the same pattern. (dynamic,* EVENT*)*
... das gleiche Muster zeigten. (stative, STATE*)*

We also identified a number of aligned clauses that received different annotations caused by a different habituality status. This is a borderline cases since this kind of difference appears repeatedly but not consistently. Most of these segments use the same tense and the same word choice so that it is hard to identify what causes this shift. None of the two languages has a higher tendency to marking clauses as habitual and the clauses in this class are equally distributed between the two languages. We give an example of this in (65).

(65) *At intervals, while turning over the leaves of my book ... (*GENERALIZING*)*
In kurzen Zwischenräumen, wenn ich die Blätter meines Buches wendete ...
*(*EVENT*)*

Annotation errors and disagreements between annotators make up round 20% of all differences. (66) is an example of an annotation error on the German part, as the type should coerce to STATE due to the negation. Annotator disagreements contribute the most in this class of mismatches as we identified 252 such cases. They often occur in clauses containing verbs with ambiguous aspectual class or habituality status. Most of these disagreements are systematic and the most confusions can be categorized into three main categories, namely the distinctions between STATE, GENERIC SENTENCE and GENERALIZING SENTENCE, between EVENT, GENERIC SENTENCE and STATEas well as between EVENT, GENERIC SENTENCE and GENERALIZING SENTENCE. In (67) we give an example of one such disagreement.

(66) *As she got no answer to this ... (*STATE*)*
*Als sie keine Antwort bekam ... (*EVENT*)*

(67) *People who have these hallucinations ... (no gold label, annotator disagreement between* STATE, GENERIC SENTENCE *and* GENERALIZING SENTENCE*)*
*Leute, die diese Halluzinationen haben ... (*GENERIC SENTENCE*)*

As we mentioned before almost half of the identified mismatches were found to be a result of cross-linguistic differences. As one would expect many of these differences stem from segments containing perfect. Most of these cases are due to the fact

that clauses containing a verb in simple or past perfect in German can be marked with three different labels (STATE, EVENT and EVENT-PERFECT-STATE). In contrast all English clauses containing perfect are considered to be states as they focus on the post-state of the action described (Katz, 2003). (68) shows an example of this.

(68) *Also I had drawn parallels in silence ...* (STATE)
Im Stillen hatte ich Vergleiche gezogen ... (EVENT)

Additional language-pair dependent mismatches result from different lexical choice and grammatical structures between the two languages. Since the aim of translation is to convey the meaning of the original using words and structures that sound natural in the target language, these kind of differences are very common and fully justified in parallel, human translated texts. Questions, for instance, may be indirect in one language and translated as direct questions in the other, as in (69). (70) serves as an example of the use of not direct translation equivalents to convey the same meaning.

(69) *I would like to know what technical measures have been taken in the new buildings in Brussels and Strasbourg.* (STATE)
Welche technischen Maßnahmen sind in den Neubauten in Brüssel und in Straßburg ergriffen worden? (QUESTION)

(70) *So a group of babies came in ...* (EVENT)
Also hatten wir auch eine Gruppe von Babies ... (STATE)

Finally, more than 15% of the total mismatches are classified as segmentation errors. This is true for clauses that do not contain full verb constellations like in the case of segments containing only infinitives and, thus, no situation at all. It is worth mentioning that more than 60% of these cases stem from Smultron which is aligned semi-automatically. An example is shown in (71).

(71) *To be sure ... (no situation)*
Trotzdem kann man sicher sein ... (STATE)

5.1.3 Event-Perfect-State: Analysis and Impact

We hypothesize that German clauses containing present or past perfect can have a stative or an event reading, or be underspecified depending on the context. For the annotation of the German side of our corpus we have introduced a new label, the

Event-Perfect-State. To support our hypothesis of the three different interpretations of the perfect, we conduct a large-scale experiment, where annotators decide which reading is more important in selected clauses that we have labeled as either State, Event or Event-Perfect-State beforehand[6]. The results validate our decision as the annotators' ratings match our labels in most cases and most importantly correlate to each other's decisions (Mavridou et al., 2015).

The newly introduced label appears 71 times in our German data as gold standard. All three annotators agree on the label in 16 segments which is round 20% of all cases. Some annotators use the label more often than others but generally it is scarcely used. Table 9 shows the various pairings of labels between annotators on all cases where at least one annotator uses the label. The most frequent pairing Ev-p-s, Event, Event as well as cases where all three annotators disagree on the three possibilities for labeling the perfect, as we see in (72). We consider this type of disagreement as language-pair dependent, belonging to the "perfect" class of the shifts.

(72) *But the philosopher had stopped her. (State)*
Aber jetzt hatte der Philosoph sie zurückgehalten. (no gold label, annotator disagreement between State, Event and Event-Perfect-State)

Pairings of labels	**Frequency**
Ev-p-s, Ev-p-s, Ev-p-s	16
Ev-p-s, Ev-p-s, State	15
Ev-p-s, Ev-p-s, Event	39
Ev-p-s, Ev-p-s, Other	1
Ev-p-s, State, State	19
Ev-p-s, State, Event	42
Ev-p-s, Event, Event	52
Ev-p-s, Event, Other	6
Ev-p-s, State, Other	4
Ev-p-s, Other, Other	9

Table 9: Frequencies of pairings of labels in segments containing at least one Event-Perfect-State annotation

[6]This experiment was ran by Melissa Peate Sørensen and Annemarie Friedrich

We calculate the inter-annotator agreement for this label by setting all other labels to "NONE". The agreement is 0.49 for set 1 without Smultron, 0.26 for Smultron (both in Fleiss' κ) and only 0.13 for set 2 (in Cohen's κ). We further split set 2 into a section containing TED talks only and another section for the rest of the documents in order to find out the reason of the low agreement for set 2. The agreement for all documents without TED talks is 0.19 while the agreement for the TED talks is only 0.08. This reflects the difficulty of annotating this specific corpus section and explains the low inter-annotator agreement on set 2.

The results suggest that the label should be further analyzed to develop clearer and more precise guidelines for its annotation. We give no specific rules and ask annotators to decide mainly based on their intuition. This certainly allows for improvement since the experiment we conducted implies that German native speakers do indeed identify three different readings. The analysis of the disagreements between the three possible labels for the perfect might prove useful in identifying rules that will make the distinction easier.

5.2 Discourse Modes - Pilot Study

The motivation of this work is that a study at the level of situation entities can lay the foundation for the study of discourse modes and discourse in general. Discourse modes, which are linguistic properties of text passages, are the key idea in the work of Smith (2003), which we briefly discussed Section 2. Smith (2003) suggests that DMs and SEs are two interacting levels of linguistic analysis. Each DM has a different distribution of SEs. So, Narratives, Report and Description have a high proportion of States and Events whereas Information and Argument introduce in addition many General Statives.

We conduct a a small pilot study by annotating paragraphs with their DM type[7]. The same texts are annotated with SE information by different annotators in order to explore the link between DMs and SEs. The little amount of available data and the fact that States are dominant in each mode do not allow to draw confident conclusions. Figure 3 and Figure 4 show the distributions of the SE types per DM in English and German, respectively, which generally match the predictions of Smith (2003). We see, for instance, that the proportion of States and Events is higher in Narrative, Report and Description, while General Statives are more frequent in the Information and Argument mode. The only exception is the higher proportion

[7]This study was ran by Alexis Palmer

of Generalizing Sentences in the Description mode in the German part, which can be considered an annotation problem, since the guidelines for the DM annotation were very short and annotators did not receive extensive training.

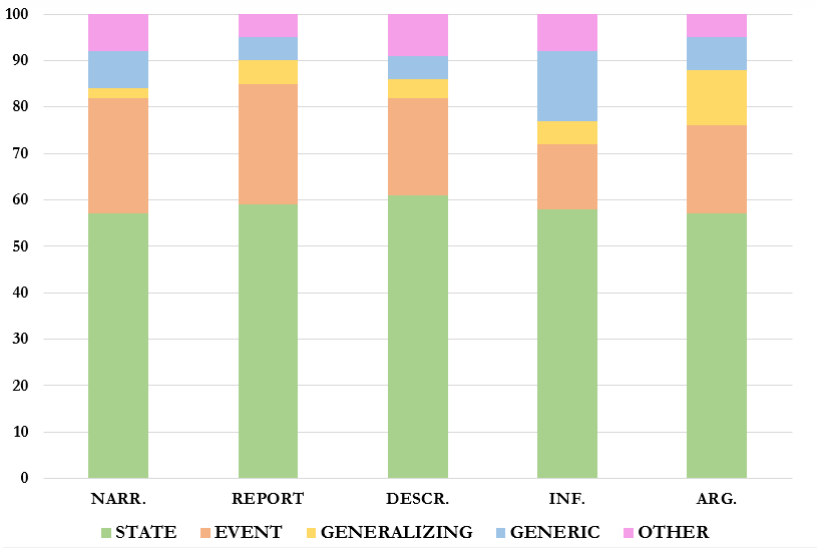

Figure 3: Distribution of SE types per DM in English.

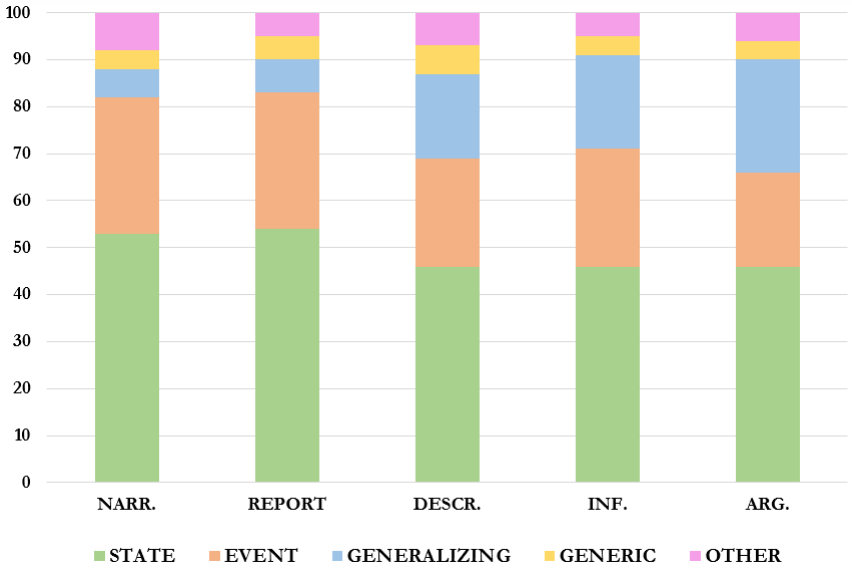

Figure 4: Distribution of SE types per DM in German.

5.3 Summary

We conduct a corpus-based cross-linguistic study to investigate whether SE types correspond across parallel, aligned data and in which cases the SE types differ. Our findings show that the majority of the aligned clauses have the same label and that the mismatches are systematic. The proportion of language-independent and language-pair dependent shifts is almost equally high. Different lexical choice and grammatical structures, clauses containing perfect, differences during the annotation of habituality and segmentation errors are types of shifts than depend on our specific language-pair. Most of the language-independent confusions that could also occur in a monolingual setting involve the labeling of the main referent as generic or non-generic. The biggest cause of this type of mismatches are annotation errors, as well as disagreements between all three annotators.

For future work we suggest looking at different language pairs, preferably not as closely related as English and German. Moreover, it would be interesting to repeat this study not on translated but on original, comparable texts. Since we confirmed that there is a link between discourse modes and situation entities it should be possible to compare texts based on their DM distributions and analyze them with respect to their cross-linguistic correspondence of SE types. Another direction for work in the future would be the automatic prediction of the SE type shifts, which could provide a valuable resource for (machine) translation studies.

6 Automatic classification of situation entity types

The second contribution of this work is the creation of the first automatic classifier for situation entity types for German, and a comparison to an already existing, reimplemented classifier for English, using basic features like part-of-speech (POS) tags, lemma and word information. To evaluate our classifiers we run our experiments using 10-fold as well as document-wise leave-one-out cross-validation. We use different feature combinations and compare our results against a baseline that is determined by assigning the label of the most common class to all clauses. The creation of such classifiers will be a useful tool to distinguish the discourse modes present in documents, which in its turn can be beneficial in discourse analysis.

Our work is inspired by and based on the work of Palmer et al. (2007). In their work, three different feature sets are used: a basic set consisting of words and part-of-speech tags, a linguistically-motivated set containing cues that correlate to certain SE types, like the presence of modal verbs or generic predicates in the clause, as well as a set of deep features extracted from CCG parses. In this thesis, we only use some variant of the basic feature set to create our classification models.

6.1 Features

The basic features of Palmer et al. (2007) consist of two feature sets: the "words" feature set that only looks at words and punctuation in the clause and is obtained without preprocessing, and the "words/tags" feature set, which includes the POS tags for each word, number and punctuation in the clause as well as the word/tag pairs for each token in the clause. We use five different feature sets, namely Tags, Lemmata, Words, as well as the combined feature sets Tags/Lemmata and Tags/-Words. Unlike the previous work, we choose lemmata as additional features because they provide valuable information like words, while at the same time reduce sparsity to some extent. Words and lemmata represent lexico-semantic information whereas the POS tags provide syntactic as well as semantic information, like the presence of verbs, nouns or the identification of verb tense.

The features for both languages are extracted using TreeTagger (Schmid, 1994), a freely available tool for annotating texts with part-of-speech and lemma information. The tool uses the PTB (Santorini, 1990) tagset for English and the STTS tagset (Schiller et al., 1999) for German.

We take some preprocessing steps in order to get our features. We generally consider words or lemmata and some informative punctuation marks, like period, comma, question mark, exclamation mark and quotes and ignore other marks and special characters, like "&", "+", "@", etc. In the case of quotation marks we replace all possible marks used for quoting by single quotes. Moreover, to further reduce sparsity we lower-case the English words and lemmata. We do not take this step in the German part, because case-sensitivity is more meaningful here, for example in the case of nominalised verbs. In addition, we rewrite contractions to their full forms (for example "don't" to "do not") as the TreeTagger often fails to correctly tag them. The number of attributes per feature set and per language after our preprocessing is shown in Table 10.

	# English attr.	# German attr.
Tags	48	50
Lemmata	5497	7642
Words	7011	10622
Tags/Lemmata	5545	7692
Tags/Words	7059	10672

Table 10: Number of attributes per feature set, per language.

6.2 Classification algorithm

In our experiments we use a Random Forest classifier trained with the implementation and the standard parameter settings from Weka (Hall et al., 2009). Random forests (Breiman, 2001) work as a large collection of de-correlated decision trees, though the approach has an advantage over decision trees: since it is an ensemble method it averages over individual trees with high variance that overfit to their training set, increasing this way the performance.

Random forests is a simple algorithm that is widely used because they are fast, simple to train, easy to interpret and prevent overfitting to the training data. The way random forests work is the following: the whole set of training examples is randomly split in K different sets. For each set of the training examples the method learns the full tree. The main idea, though, is that this is done only for a subset of examples using a subset of attributes. It is worth noting,that each of the generated trees considers

different attributes. At prediction time, new data points are classified using each of the generated trees and the class is predicted by majority vote.

6.3 Evaluation Methodology

The commonly used holdout evaluation method that splits the data into a training, a development and a test set is not the best option in our case because our corpus is small and the splitting might cause partitions with different distributions, due to the variety of genres in the corpus. Instead, we are running our experiments using stratified 10-fold and leave-one-out cross-validation.

In the stratified *k*-fold cross-validation, which Weka (Hall et al., 2009) uses by default, the original data set is randomly partitioned into *k* equal size subsamples. In contrast to the simple cross-validation, however, the folds in the stratified cross-validation are selected so that the distribution of the class labels in the entire corpus is maintained in each fold. In each round, one of the *k* subsets is used as test set and the other *k*-1 subsets are used as training set, and the process is repeated *k* times. The results from the *k* folds can then be averaged to give a single estimation. The advantage of this method over the conventional validation that splits the original data set into two or three parts is that all observations are used for both training and testing, and each observation is used for testing only once, reducing at the same time the variance in the estimate.

In stratified cross-validation the situation entities of a document are distributed across the folds and can be used both for training and testing in a single fold. In the document-wise leave-one-out cross-validation, on the other hand, in each fold all but one documents are used for training and one document is reserved for testing. The approach maximizes, thus, the use of the data which is important in our case where only a small amount of data are available. The main disadvantage of this approach is that it cannot be stratified, which means that the different folds do not reflect the overall distribution of classes in the corpus

6.4 Evaluation Results

Our results are reported in terms of accuracy, which is the percentage of correctly labeled clauses, as well as in terms of precision, recall and F-measure. The baseline is determined by assigning the label of the most frequent class (STATE) to each clause.

The English baseline is higher than the German because the proportion of states in the data is higher. This is probably caused mainly by the perfect coercion, which on the German part is not always necessary. As mentioned before, a German clause containing perfect can be assigned the label STATE, EVENT or EVENT-PERF-STATE, whereas in English it always gets labeled as STATE.

Table 11 and Table 12 show the results of the 10-fold and 30-fold cross-validation on the English data. In the 10-fold cross-validation setting, the use of part-of-speech information alone increases accuracy by almost 10% and F-measure from 0.37 to 0.61. The use of LEMMATA and WORDS further increases accuracy to around 65% against 53.5% of the baseline and F-measure to 0.65 and 0.66, respectively. The best results for all four measures (accuracy, precision, recall and F-measure) are achieved using the feature set TAGS/LEMMATA. This boosts accuracy to 66.1% and F-measure to 0.68, followed by the combination of TAGS/WORDS with 65.3% accuracy and 0.66 F-measure.

We would expect the leave-one-out cross-validation to have similar (though lower) results but the pattern is different in this setting. TAGS/LEMMATA is the feature set with the best results in terms of accuracy in this case as well, though at the same level as the TAGS alone with 63.5% accuracy. TAGS/WORDS has the highest precision score (0.74) which boosts F-measure to 0.68, the best among all feature sets, followed by WORDS with 0.65. Unlike in the 10-fold cross-validation the results seem to be more random, with TAGS and TAGS/LEMMATA achieving highest accuracy scores while WORDS and TAGS/WORDS being the two best sets in terms of F-measure.

Generally, the highest differences in accuracy between the two settings are less than 3 points, while the same F-measure is reached in both cases, though using different feature sets. The highest gain in accuracy using 10-fold cross-validation is 12.6%, while the best accuracy in the 30-fold cross-validation setting beats the baseline accuracy by 10%. In both settings, the best F-measure is 0.68 or 0.31 points higher than the baseline.

Although we use different data sets and models for the classification of situation entities, our results are comparable with those of Palmer et al. (2007). In their work, Palmer et al. (2007) use 20 documents from the "popular lore" section of the Brown corpus which consist of 4390 clauses and span a wide range of topics and situation entity types. Our data set on the other hand is slightly bigger, containing 5783 English clauses, covering different genres as well. In their work, a maximum entropy model is used to predict the situation entity types of clauses, while the standard 10-fold cross-validation is used to develop the model. Their baseline accuracy is 38.5%. They

do not report on the gain from adding POS information alone, however, by adding words their model beats the baseline by 6.9%. By adding tags to the words they achieve an increase of 11.4%, which is almost the same as the increase of 11.6% we achieve by using our Tags/Words feature set.

	Accuracy	Precision	Recall	F-measure
baseline	53.5	0.29	0.53	0.37
Tags	63.3	0.59	0.63	0.61
Lemmata	65.0	0.65	0.65	0.65
Words	65.1	0.68	0.65	0.66
Tags/Lemmata	**66.1**	0.70	0.66	**0.68**
Tags/Words	65.3	0.68	0.65	0.66

Table 11: 10-fold cross-validation results for English.

	Accuracy	Precision	Recall	F-measure
baseline	53.5	0.29	0.53	0.37
Tags	**63.5**	0.64	0.63	0.64
Lemmata	62.2	0.60	0.62	0.61
Words	63.2	0.66	0.63	0.65
Tags/Lemmata	**63.5**	0.64	0.63	0.64
Tags/Words	63.2	0.74	0.63	**0.68**

Table 12: 30-fold cross-validation results for English.

A similar increase against the baseline is also achieved on the German part. Table 13 shows the results for the 10-fold cross-validation. The simple Tags set beats the baseline by 7.6% with 52.1% accuracy. Words is the feature set with the best accuracy of 59.0% and is 14.5% higher than the baseline, a gain higher than that on the English part. In terms of F-measure, Lemmata has the best performance with 0.58 (0.31 higher than the baseline), followed by Words and Tags/Words with 0.57.

	Accuracy	Precision	Recall	F-measure
baseline	44.5	0.20	0.45	0.27
Tags	52.1	0.49	0.52	0.50
Lemmata	57.6	0.59	0.58	**0.58**
Words	**59.0**	0.55	0.59	0.57
Tags/Lemmata	58.6	0.53	0.59	0.56
Tags/Words	58.2	0.57	0.58	0.57

Table 13: 10-fold cross-validation results for German.

	Accuracy	Precision	Recall	F-measure
baseline	44.5	0.20	0.45	0.27
Tags	49.4	0.36	0.49	0.42
Lemmata	52.1	0.41	0.52	0.46
Words	53.1	0.50	0.53	0.52
Tags/Lemmata	53.5	0.49	0.54	0.51
Tags/Words	**53.9**	0.52	0.54	**0.53**

Table 14: 30-fold cross-validation results for German.

In the 30-fold cross-validation setting, the gains over the baseline are again lower, as expected. Still, the use of Tags alone increases accuracy by 4.9%. The use of the other feature sets further increases accuracy with Tags/Words being the set with the highest accuracy 53.9%, which translates to a gain of 9.4% against the baseline, comparable but slightly lower than the best performance on the English part. Tags/-Words shows the best performance not only in terms of accuracy but also has the best F-measure 0.53, followed by Words with an F-measure of 0.52.

6.4.1 Discussion

Overall, the results are as we expected. There are a few things, however, that we find interesting. First of all, the best feature sets differ across the two validation settings for both languages. In the German part, Words and Lemmata are the best-performing

feature sets in the 10-fold cross-validation in terms of accuracy and F-measure, respectively, whereas Tags/Words performs best in the 30-fold cross-validation setting. However, since we think that the leave-on-out cross-validation is more realistic, the observations we report on are focused on this setting.

Moreover the best feature sets do not differ only across the two validation settings but also cross-linguistically. Tags/Lemmata has the highest performance gain in terms of accuracy when tested on the English data, whereas Tags/Words is the best feature set on the German data, though the results are very close. A possible explanation for this is that in German there are more words contained in the feature set, which are more informative than the English, like for example the past participle which has a distinct form and does not use the simple past form. In the English part, on the other hand, lemmata are providing useful information and at the same time account to some point for sparsity issues in the data.

Another interesting observation is that part-of-speech information alone seems to be less informative in German. While the use of the Tags feature set improves performance by almost 10% in the English data, the equivalent gain on the German part is less than 5%. This can be due to the fact that English has finer-grained tags than German, for instance different labels for verbs in present and past tense. In fact, we have created a manual mapping between the German STTS and English PTB tagsets and found cases where eight distinct or five distinct English verb tags correspond to a single German tag. Moreover, English has a distinct POS tag for nouns in plural, something that German lacks, which might be useful in the distinction between generic and non-generic main referents.

In order to prepare for our next experiment, the domain adaptation experiment described in the next Section, we run some additional experiments using UTS, a tagset that consists of twelve universal part-of-speech categories (Petrov et al., 2011). For these experiments we replace our tagsets with the UTS tagset using the mapping created by the authors[8]. The tags used in this tagset are shown in Table 15.

Tables 15-16 show the results of our experiment when using UTS. We have experimented with Tags and Tags/Lemmata and did not consider Tags/Words, as we did not expect any improvement using this feature set either. The results using UTS-Tags/Lemmata are worse than using any other feature set except for Tags only. When using the UTS-Tags alone the results are even worse than the baseline. Thus, it is clear that the fine-grained POS categories are contributing a lot to our work.

[8]https://github.com/slavpetrov/universal-pos-tags

VERB	verbs (all tenses and modes)
NOUN	nouns (common and proper)
PRON	pronouns
ADJ	adjectives
ADV	adverbs
ADP	adpositions (prepositions and postpositions)
CONJ	conjunctions
DET	determiners
NUM	cardinal numbers
PRT	particles or other function words
X	other: foreign words, typos, abbreviations
.	punctuation

Table 15: The Universal Part-of-Speech Tagset (UTS).

	Accuracy	Precision	Recall	F-measure
baseline EN	53.5	0.29	0.53	0.37
TAGS EN	63.3	0.59	0.63	0.61
UTS-TAGS EN	49.7	0.43	0.50	0.46
TAGS/LEMMATA EN	66.1	0.70	0.66	0.68
UTS-TAGS/LEMMATA EN	62.7	0.64	0.63	0.63
baseline DE	44.5	0.20	0.45	0.27
TAGS DE	52.1	0.49	0.52	0.50
UTS-TAGS DE	39.9	0.35	0.40	0.37
TAGS/LEMMATA DE	58.6	0.53	0.59	0.56
UTS-TAGS/LEMMATA DE	55.7	0.56	0.56	0.56

Table 16: 10-fold cross-validation results for English and German using UTS.

	Accuracy	Precision	Recall	F-measure
baseline EN	53.5	0.29	0.53	0.37
TAGS	63.5	0.64	0.63	0.64
UTS-TAGS EN	48.8	0.28	0.49	0.36
TAGS/LEMMATA	63.5	0.64	0.63	0.64
UTS-TAGS/LEMMATA EN	60.28	0.64	0.60	0.62
baseline DE	44.5	0.20	0.45	0.27
TAGS DE	49.4	0.36	0.49	0.42
UTS-TAGS DE	39.3	0.22	0.39	0.29
TAGS/LEMMATA DE	53.5	0.49	0.54	0.51
UTS-TAGS/LEMMATA DE	51.4	0.45	0.51	0.48

Table 17: 30-fold cross-validation results for English and German using UTS.

6.5 Domain Adaptation Experiment

In order to improve our classification model for German we need a big amount of annotated data. In our case our German data are sparse but we do have a bigger amount of English annotated data. Daumé III (2007) suggests an innovative and simple approach to domain adaptation that is suitable exactly in cases where more annotated "target" data are available. We assume the different languages to be different domains and experiment with this approach using two additional English data sets as additional training data.

6.5.1 Approach

The idea of Daumé III (2007) is very simple. Essentially, what this approach does is taking each feature in the source side and creating three versions out of it: a general version, a source-specific version and a target-specific version.

To put this formally, X and Y are the input and output spaces respectively, and $\Phi^s, \Phi^t : X \rightarrow \check{X}$ the mapping for source and target data, respectively. The augmented source data contains only general and source-specific versions of the features, whereas

the augmented target data contains only general and target-specific versions, or formally:

$$\Phi^s(x) = \langle x,x,0 \rangle, \quad \Phi^t(x) = \langle x,0,x \rangle$$

A supervised classifier is then trained on the features from both domains. The approach delivers solid performance on a variety of domains and tasks, like part-of-speech tagging, named-entity recognition and shallow parsing. The success lies mainly in the general version of the features that is able to capture common, domain-independent features.

We try out this approach considering different languages to be different domains. We run two experiments using the TAGS feature set only. For the purpose of the experiment we map English to German tags to create a common tagset between our two languages. We then train our classifier on three data sets in addition to our German data: our parallel English data set, as well as the WikiGenerics corpus[9] (Friedrich et al., 2015) and the MASC[10] corpus. WikiGenerics consists of 102 documents extracted from Wikipedia containing a high proportion of generic sentences, like documents about animals, instruments, games, etc. The MASC corpus, on the other hand, consists of 205 documents covering a wide range of genres like blogs, fiction, letters and news texts. We develop our model using leave-one-out cross-validation, training at each fold on all but one German documents plus the English documents of the respective data set.

As mentioned before, we create a mapping between English and German POS tags and replace English with German tags so that in the end we only use labels from the STTS tagset. However, since the English POS categories are more fine-grained, we end up with several many-to-one mappings. This is true, for instance, for several verb, adjective or adverb tags that correspond to a single tag in German. Due to this mapping we lose useful information that the finer-grained English categories carry but since we want to improve our German classification model this does not pose a big problem.

6.5.2 Results

We first run the domain adaptation experiment using our parallel English corpus as out-of-domain data. Our results beat the baseline and are slightly better than

[9] http://www.coli.uni-saarland.de/projects/sitent/data/WikiGenerics%20v2.0.zip

[10] to be published soon at http://www.coli.uni-saarland.de/projects/sitent/page.php?id=resources

our results using German data only, though only in terms of F-measure. Since the German and English data are parallel, using the English part as additional data for training our model might not add new information and, thus, not change the results. To overcome this, we experiment with the WikiGenerics corpus as additional training data. The results remain almost unchanged with regard to the previous experiment and are even worse in terms of accuracy. A possible explanation for this might be the fact that the WikiGenerics data have a different distribution of POS tags and are therefore not of much help for our task. As a last experiment we use the MASC data that consist of different genres and have a lower proportion of generic sentences. The results of this experiment are better than both the baseline and our German results, although the improvement is very small.

We assume that the different POS tag patterns present in the two languages might not help in predicting the type of situation entity of a clause. Although they are closely related languages, in many cases the POS patterns differ due to the different grammatical or lexical constructions used in each language or the opposite: same POS patterns occur for signaling different situation entity types, like in the example of perfect, that always introduces a state in English but not necessarily in German. Table 18 summarizes the results of the above-mentioned experiments.

	Accuracy	Precision	Recall	F-measure
baseline	44.5	0.20	0.45	0.27
TAGS DE	49.4	0.36	0.49	0.42
TAGS & parallel EN	49.3	0.36	0.49	0.42
TAGS & WIKI	49.1	0.36	0.49	0.42
TAGS & MASC	49.5	0.41	0.50	0.45

Table 18: Domain adaptation results for German using our parallel English data, WikiGenerics and MASC as additional training data.

We had hoped that this approach would deliver more solid results. In order to test whether the non-improving results are due to the cross-linguistic nature of the task, we repeat the experiments testing on the English side of our data. This time, no mapping is needed so we eliminate the chance of information loss due to less informative POS categories. The results, shown in Table 19, are better than the baseline but worse than those we get using our English data set only. Although the suggested approach delivers promising results for different tasks, we do not see the benefits in our work.

	Accuracy	Precision	Recall	F-measure
baseline	53.5	0.29	0.53	0.37
TAGS	63.5	0.64	0.63	0.64
TAGS & WIKI	61.0	0.48	0.61	0.54
TAGS & MASC	61.6	0.44	0.62	0.51

Table 19: In domain results for English using WikiGenerics and MASC as additional training data.

6.6 Summary

Based on previous work (Palmer et al., 2007), we build an automatic classifier for labeling German clauses with their situation entity type. For comparison we reimplement and extend an already existing classifier for English. We use five different simple feature sets consisting of part-of-speech, word and lemma information and combinations of POS with words and lemmata. Lemmata seem to contribute the most in the English part, while words are the most informative features for German. The highest accuracy gain we achieve against the baseline is 9.4% and 10% for German and English, respectively, evaluating with leave-one-out cross-validation. The results are higher (14% and 12.6% gain, respectively) when using 10-fold cross-validation, and comparable to the results of Palmer et al. (2007).

Due to the small amount of annotated German data we explore the domain adaptation approach suggested by Daumé III (2007). We consider our English data as out-of-domain and expand the feature space by creating three versions of the features: a domain-independent, a source-specific and a target-specific version. As a first step we run two experiments using UTS, a universal tagset consisting of twelve tags. This reduction in the number of POS tags results in a worsening in the performance than when using the whole tag inventory, especially for German. Since this mapping is too coarse-grained to deliver solid results for our domain adaptation experiment we create a manual mapping between English and German tags. We use two bigger, annotated English data sets as training data in addition to our German data. The performance only slightly increases when we train on MASC, a data set consisting of documents from similar domains as our corpus.

7 Conclusion

7.1 Results

Situation entities are well-studied in linguistics but only a few works address their computational processing. Most related work focus on certain types, like states, events or generics, mostly separately. In this work, we have first studied whether the situation entity types of parallel texts correspond across two languages. Our second contribution is the creation of two classifiers, an English one, based on an already implemented model, as well as a German one. To our best knowledge this is the first classifier built to label clauses of German texts with their situation entity types.

As a first step we adjusted the English annotation scheme of Friedrich and Palmer (2014b) in order to apply it on German data. The guidelines seem to be clear to annotators as agreement figures are at the same levels as for the English part. After this, we created the first English-German parallel corpus annotated with situation entities. Following our observations we introduced a new situation entity type, the Event-Perfect-State, used for clauses containing perfect, that are underspecified and cannot be categorized neither as States nor as Events. This observation is supported by a large-scale web experiment (Mavridou et al., 2015), where annotators were asked to rate how eventive or how stative the reading of several selected clauses was. The annotators' ratings match our reference ratings in most cases.

With the use of our parallel corpus we conducted a cross-linguistic study to check whether the situation entity types are the same in aligned clauses. Our findings show that most instances correspond cross-linguistically. We extracted all differences and analyzed them thoroughly. What we found is that half of them are a result of cross-linguistic differences, like clauses containing perfect or clauses that convey the same meaning using different lexical items or grammatical constructions. The rest of the disagreements could occur in a monolingual setting as well and they mainly involve the genericity of the main referent.

After the corpus study we created two classification models for automatically labeling English and German clauses with their SE types. Based on previous work (Palmer et al., 2007) we use five different feature sets containing part-of-speech, lemma and word information, as well as combinations thereof. Our best results beat the baseline by 12.6% in the English part and by 14% in the German part, with an accuracy of 66% and 59%, respectively. Tags/Lemmata works best for English, while Words and Tags/Words are most informative for the German classifier.

Due to the small amount of German annotated data we also experimented with the domain adaptation approach suggested by Daumé III (2007). In this scenario, annotated data from a different domain are used as additional training data and the feature space is expanded by creating three versions of the features: a domain-independent, a source-specific and a target-specific version. We tested this approach by using two additional English data sets as training data, after having mapped English POS tags to German POS tags to create common ground for the experiment. The impact is only minor using TAGS only and training on one of the data sets that has a similar SE types distribution as our corpus, which leads us to the conclusion that the addition of more information is needed to achieve better results.

7.2 Future Work

There are parts in our work that allow for improvement. First of all, the annotation manual for German can be expanded in many ways. For this thesis we based our adjustments and suggestions on observations from the data. Other data sets might reveal the need for further development. Moreover, some of our choices are based on intuition, like the newly introduced Event-Perfect-State type. These suggestions need to be thoroughly analyzed within a formal framework to reduce confusion, which in its turn could possibly improve inter-annotator agreement.

We believe that it would be worth to investigate whether the mismatches we identified manually in our corpus study can be predicted with an automatic classifier. Such a resource would be a valuable tool for translation studies or even for evaluating machine translation output.

Our German classification model can also be improved in many ways. The first obvious idea is to use semantic-syntactic features which are already being used for English and investigate if they prove equally informative for German. Moreover, we only had a small corpus of German annotated data. It would be interesting to test whether performance would rise by adding more data from various domains. Although we tried to have a fair amount of generics in our corpus, eventualities were dominant in the texts. It would make sense to experiment with data with a more balanced distribution of situation entities, as this is something that could also affect the performance.

References

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy*, 9(1):5–16.

Bethard, S. and Martin, J. H. (2006). Identification of event mentions and their semantic class. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 146–154. Association for Computational Linguistics.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Bihl, J. (1949). *German One: A Cultural Approach*. Houghton Mifflin Company.

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

Carlson, G. N. (2005). Generics, habituals and iteratives. In Barber, A., editor, *Encyclopedia of Language and Linguistics*. Elsevier.

Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Friedrich, A. and Palmer, A. (2014a). Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Baltimore, USA.

Friedrich, A. and Palmer, A. (2014b). Situation entity annotation. *LAW VIII*, page 149.

Friedrich, A., Palmer, A., Sorenson, M., and Pinkal, M. (2015). Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop (LAW IX)*.

Friedrich, A. and Pinkal, M. (2015a). Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481, Lisbon, Portugal. Association for Computational Linguistics.

Friedrich, A. and Pinkal, M. (2015b). Discourse-sensitive automatic identification of generic expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U., and de Jong, F. (2011). Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1061–1070, New York, NY, USA. ACM.

Helbig, G. and Buscha, J. (2001). Deutsche grammatik: ein handbuch fur den auslanderunterricht.

Hentschel, E. (2003). Es war einmal ein subjekt. *Linguistik online*, 13:137–160.

Hongye, T., Tiejun, Z., and Jiaheng, Z. (2008). Identification of chinese event and their argument roles. In *Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on*, pages 14–19.

Islam, Z. and Mehler, A. (2012). Customization of the europarl corpus for translation studies. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Katz, G. (2003). On the stativity of the english perfect. In *Perfect explorations, de Gruyter, The Hague.205233*.

Kenny, A. (1963). *Action, Emotion and Will*. Studies in philosophical psychology. Routledge.

Klein, D. and Manning, C. D. (2003). Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Klein, W. (2000). An analysis of the german perfekt. *Language*, 76(2):358–382.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Krifka, M., Pelletier, F., Carlson, G., Chierchia, G., Link, G., and Ter Meulen, A. (1995). Introduction to genericity. *The generic book*, pages 1–124.

Louis, A., Joshi, A., and Nenkova, A. (2010). Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 147–156, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marovic, M., Šnajder, J., and Glavaš, G. (2012). Event and temporal relation extraction from croatian newspaper texts. In *Proc. of the Eighth Language Technologies Conference. Slovenian Language Technologies Society*.

Mathew, T. A. and Katz, E. G. (2009). Supervised categorization for habitual versus episodic sentences.

Mavridou, K.-I., Friedrich, A., Sørensen, M. P., Palmer, A., and Pinkal, M. (2015). Linking discourse modes and situation entity types in a cross-linguistic corpus study. In *Proceedings of Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*.

Meyer, T. (2014). *Discourse-level Features for Statistical Machine Translation*. PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL).

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., Meyers, A., Brunstein, A., Ferro, L., and Sundheim, B. (2003). Ace-2 version 1.0.

Palmer, A. and Friedrich, A. (2014). Genre distinctions and discourse modes: Text types differ in their situation type distributions. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and NLP,*.

Palmer, A., Kuhn, J., and Smith, C. (2004). Utilization of multiple language resources for robust grammar-based tense and aspect classification. In *Proceedings of LREC 2004*.

Palmer, A., Ponvert, E., Baldridge, J., and Smith, C. (2007). A sequence model for situation entity classification. In *In Annual Meeting of the Association of Computational Linguistics (ACL)*.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. In *IN ARXIV:1104.2086*.

Prasad, R., Joshi, A., Dinesh, N., Lee, A., Miltsakaki, E., and Webber, B. (2005). The penn discourse treebank as a resource for natural language generation. In *In Proc. of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Reiter, N. and Frank, A. (2010). Identifying Generic Noun Phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 40–49, Uppsala, Sweden. Association for Computational Linguistics.

Ryle, G. (1949). *The Concept of Mind*. Hutchinson.

Santorini, B. (1990). Part-Of-Speech tagging guidelines for the Penn Treebank project (3rd revision, 2nd printing). Technical report, Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA.

Saurí, R., Knippen, R., Verhagen, M., and Pustejovsky, J. (2005). Evita: A robust event recognizer for qa systems. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 700–707, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Schiller, A., Teufel, S., and Stockert, C. (1999). Guidelines fur das tagging deutscher textcorpora mit stts.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Schmid, H. (2004). Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *Proceedings of Coling 2004*, pages 162–168, Geneva, Switzerland. COLING.

Siegel, E. V. and McKeown, K. R. (2000). Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Comput. Linguist.*, 26(4):595–628.

Smith, C. S. (1991). *The parameter of aspect*. Kluwer Academic Publishers, Dordrecht.

Smith, C. S. (2003). *Modes of discourse: The local structure of texts*. Cambridge University Press.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

UzZaman, N. and Allen, J. F. (2010). Event and temporal expression extraction from raw text: First step towards a temporally aware system. *International Journal of Semantic Computing*, 4(04):487–508.

Vendler, Z. (1957). Verbs and times. *Philosophical Review*, 66(2):143–160.

Vinay, J., Darbelnet, J., Sager, J., and Hamel, M. (1995). *Comparative Stylistics of French and English: A Methodology for Translation*. Benjamins translation library. John Benjamins Publishing.

Volk, M., Göhring, A., Rios, A., Marek, T., and Samuelsson, Y. (2015). Smultron (version 4.0) — the stockholm multilingual parallel treebank. An English-French-German-Quechua-Spanish-Swedish parallel treebank with sub-sentential alignments.

Xu, H. and Huang, C.-R. (2014). Annotate and identify modalities, speech acts and finer-grained event types in chinese text. *Workshop on Lexical and Grammatical Resources for Language Processing*, page 157.

Xue, N. and Zhang, Y. (2014). Buy one get one free: Distant annotation of chinese tense, event type and modality. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Zarcone, A. and Lenci, A. (2008). Computational models for event type classification in context. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Zhang, Y. and Xue, N. (2014). Automatic inference of the tense of chinese events using implicit linguistic information. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.