Linking discourse modes and situation entity types in a cross-linguistic corpus study

Kleio-Isidora Mavridou[1] **Annemarie Friedrich**[1] **Melissa Peate Sørensen**[1]
Alexis Palmer[2,3] **Manfred Pinkal**[1]

[1]Department of Computational Linguistics, Universität des Saarlandes, Germany
{mavridou,afried,melissap,pinkal}@coli.uni-saarland.de
[2]Institute for Natural Language Processing, University of Stuttgart, Germany
[3]Department of Computational Linguistics, Heidelberg University, Germany
palmer@cl.uni-heidelberg.de

Abstract

The main contribution of this paper is a cross-linguistic empirical analysis of two interacting levels of linguistic analysis of written text: situation entity (SE) types, the semantic types of situations evoked by clauses of text, and discourse modes (DMs), a characterization of passages at the sub-document level. We adapt an existing annotation scheme for SEs in English to be used for German data, with a detailed discussion of the most important differences. We create the first parallel corpus annotated for SEs, and the first DM-annotated corpus. We find that: (a) the adapted scheme is supported by evidence from a large-scale experimental study; (b) SEs mainly correspond to each other in parallel text, and a large part of the mismatches are systematic; (c) the DM annotation task can be performed intuitively with reasonable agreement; and (d) the annotated DMs show the predicted differences in the distributions of SE types.

1 Introduction

There are complex and interwoven relationships between the nature of a text – whether construed as genre, register, text type, discourse mode, or something else – and the linguistic characteristics of the text (Werlich, 1975; Smith, 2003; Biber and Conrad, 2009; Passonneau et al., 2014, among others). Furthermore, these relationships involve phenomena at different levels, from lexical to structural, and from semantic to functional/pragmatic. In this paper we investigate correspondences across two levels of linguistic analysis, for phenomena spanning semantics and discourse, for two languages (English and German). Specifically, we conduct a corpus study on **discourse modes** (DMs), defined as types over passages of text, and **situation entity** (SE) types, defined as situation types evoked by clauses of text.

The theory of DMs (Smith, 2003) builds on the intuition that, in any genre, texts are made up of passages which have different functions. For example, a news article about student loan debt may begin with a NARRATIVE passage describing a particular student experiencing a difficult financial episode and then move on to a passage in INFORMATION mode giving background on relevant laws and policies. The different modes of discourse have different linguistic properties, one of which is the distribution of SE types predominant in the mode. (More details on DMs appear in Section 3.) We perform the first pilot annotation study of texts for DMs. Annotators label passages with their DM without referring to SEs, but only following a short manual providing prototypical examples of each DM. Our aim is to determine how easily modes can be distinguished in an intuitive setting, and to look at cross-linguistic correspondence of DM types per paragraph.

The SE types differentiate between clauses describing events, those describing states, and those conveying generic information (for more detail, see Section 4). While these semantic types are language-independent, they differ in their linguistic realizations. Here we perform the first detailed cross-linguistic study of SE types, aiming to understand both the differences in their linguistic characteristics across languages (Section 4.1) and how closely SE types correspond to each other cross-linguistically (Section 5.2). This requires adaptation of an existing annotation scheme for SEs in English (Friedrich and Palmer, 2014) to German. We discuss this adaptation (Section 4.1), including an experiment on the interpretation of

Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 12–21,
Lisboa, Portugal, 18 September 2015. ©2015 Association for Computational Linguistics.

the German perfect (Section 4.2). During the development of the annotation scheme, we identified clauses with perfect tense as one of the most difficult cases for SE annotation in German. Our corpus study shows that SE types are mostly stable across translated segments, but that there are systematic SE type shifts.

Finally, we investigate the correspondence of DMs and SEs, and find that the intuitively labeled DMs mostly have the characteristic SE type distributions predicted by Smith (2003). This is the first empirical validation of this correspondence. Interestingly, some of the pairwise DM distinctions which seem to be most difficult for annotators to make also have similar SE distributions.

In this work, we study these two related levels of semantic and discourse analysis for two reasons. The first is to provide an empirical analysis for the linguistic theory of DMs; the second is their potential to support applications like summarization, information extraction, or question answering, all of which could benefit from sorting the information conveyed by texts into different categories and different modes of presentation. Further, we discuss the potential of this level of analysis for translation studies or application within machine translation.

Related Work. Unlike genre, a notion of text type for entire documents, DMs are an aspect of sub-document structure, and thus are similar to approaches such as Argumentative Zoning (AZ) (Teufel, 2010). AZ analyzes scientific research articles according to the rhetorical functions of their text passages, identifying and labeling passages with categories like *General scientific background* or *Contrastive/comparative statements*. The key difference is that AZ is a genre-specific approach, and DMs are relevant for most written text genres.

Liakata et al. (2013) use AZ to improve summarization of scientific articles, showing that sub-document structure can indeed be useful in downstream applications. Santini (2006) also employ types over passages of text (called simply "text types"), with labels that are partially similar to Smith's DMs. These text types are then used as building blocks for automatic web genre classification.

Palmer and Friedrich (2014), inspired by Webber (2009), investigate the distribution of SE types for various genres of text. In contrast, here we study the distribution of SE types per DM. Related work for the other subparts of the study is discussed in the relevant sections of the paper.

2 Corpus Data

This study requires aligned parallel data with different text types. We collect 11 parallel English-German texts from a variety of sources and produce clause- and paragraph-level alignments for the texts. Table 1 gives statistics on the number of segments, tokens, and paragraphs in each document, as well as aggregate statistics for the corpus. The translation direction differs across documents, and part of the data consists of translations from a third language into both English and German.[1] The corpus includes three documents from a version of Europarl customized for translation studies (Islam and Mehler, 2012), two documents from the news commentary corpus (WMT 2013 shared task training data[2]), sections from the novels *Alice in Wonderland* and *Anna Karenina* from the OPUS collection (Tiedemann, 2012),[3] and two texts from a multilingual news website.[4] These texts were segmented into clauses manually by one of the authors. English and German segments were also aligned manually.

In addition, we use two documents (*Sophie's world* and *economy*) from the Smultron corpus (Volk et al., 2010). We split the English part of Smultron into clauses using SPADE (Soricut and Marcu, 2003), and the German part using a syntax-based discourse segmenter for German.[5] The Smultron corpus provides alignments on a token-/phrase-level, but these phrases do not necessarily match the clause segmentation. To align clauses, we first identify the main verb of each English segment using dependency parses (Klein and Manning, 2002). We then align each segment to the German segment containing the verb to which the identified (English) main verb is aligned. For all texts, paragraph segmentation follows the paragraph breaks in the original source texts.

3 Annotating discourse modes

This exploratory study takes the first steps toward computational treatment of DMs, resulting in the first corpus of texts labeled with DMs.

[1] This metadata is available for each document pair.
[2] `http://statmt.org/wmt13`
[3] `http://opus.lingfil.uu.se`
[4] `http://globalvoicesonline.org`
[5] Publication in preparation.

source	text/excerpt	# tokens	# aligned tokens	# clauses	# aligned clauses	# aligned paragraphs
OPUS: novels	Alice in Wonderland (en) Alice in Wonderland (de)	764 690	684 647	106 98	90	10
OPUS: novels	Anna Karenina (en) Anna Karenina (de)	592 679	543 571	83 86	73	9
Europarl	Document1 (en) Document1 (de)	551 487	454 466	59 50	47	6
Europarl	Document2 (en) Document2 (de)	1879 1662	1669 1598	192 172	163	14
Europarl	Document3 (en) Document3 (de)	923 859	774 764	104 100	85	9
GlobalVoices	Heimkino (en) Heimkino (de)	816 734	689 647	102 95	84	16
GlobalVoices	Karneval (en) Karneval (de)	1014 827	847 756	89 78	72	25
NewsCommentary	Kernspaltung (en) Kernspaltung (de)	831 849	788 727	82 89	75	17
NewsCommentary	Musharraf (en) Musharraf (de)	751 770	667 714	82 78	72	12
Smultron	Sophie's World (en) Sophie's World (de)	7011 6389	5953 6825	931 937	557	188
Smultron	Economy (en) Economy (de)	10312 9532	4238 3894	863 740	471	184
TOTAL	English German	25444 23478	17306 17609	2693 2523	1789	490

Table 1: Size of English-German parallel corpus, with per-document statistics.

3.1 Annotation scheme and analysis

Annotating DMs involves two aspects: finding the boundaries between passages of different DMs and labeling those passages with the appropriate DM. In this study we take paragraphs as an approximation of DM segments, leaving the modeling of DM boundaries for future work. The DM types used in this study are described below, together with some of the linguistic characteristics of the modes identified by Smith. These characteristics are of two types: the distribution of SE types (Section 6) and the mode of progression through the text.

- NARRATIVE: mode used for telling stories; temporal progression is generally linear
- REPORT: typical mode of news articles; events are discussed with respect to a reference time
- INFORMATION: mode used for explanations; atemporal, often focuses on generalizations rather than specific entities or events
- DESCRIPTION: mode used to describe entities, locations, objects; temporally static, progression often spatially oriented
- ARGUMENT/COMMENTARY: mode used for persuasion or presenting opinions; atemporal
- OTHER: text types not covered by Smith's set of DMs, such as instructional texts
- NONE: paragraphs whose text serves primarily structural purposes, such as headlines or document section headings

One aim of this pilot annotation is to determine how intuitively clear these categories are to minimally-trained annotators. Annotators were given a short, simple annotation manual of just 2 pages, focusing on intuitive descriptions of the modes with a prototypical paragraph for each DM. The training phase consisted of labeling and getting feedback on 14 paragraphs of text, with 2 examples of each type. The training examples were selected to be clear cases, in order to give the annotators a strong intuitive sense of each DM. Once annotators had completed the training examples, they were given documents packaged in chunks of 30 consecutive paragraphs each. Ten different annotators each labeled from 3-7 such chunks. Each paragraph is labeled once, with five annotators labeling English text, and five labeling German text.

Agreement between annotators. Inter-annotator agreement is captured through an *agreement* chunk containing five 10-paragraph segments extracted from different texts, taking aligned paragraphs for the two languages. All 10 annotators labeled the agreement dataset, five for each language. For these 50 paragraphs, Fleiss' κ for the German-language annotators is 0.50, with κ of 0.46 for the English-language annotators.

		German						
		NARR.	REPORT	INF.	DESCR.	ARG./COMM.	OTHER	NONE
English	NARRATIVE	43	5	2	8	6	1	2
	REPORT	0	17	65	4	3	0	2
	INFORMATION	9	13	53	23	10	5	3
	DESCRIPTION	16	6	8	20	7	6	3
	ARG/COMM	4	1	5	3	48	5	4
	OTHER	2	0	0	1	2	2	12
	NONE	1	1	4	1	2	10	42

Table 2: Confusion matrix of DM paragraph labels for parallel English-German text. Lightly-shaded cells highlight the most prominent confusions.

During the annotation process, it quickly became clear that two distinctions in particular were difficult for annotators to make: DESCRIPTION vs. INFORMATION, and INFORMATION vs. REPORT. Below we show three passages with their true labels. Nearly all annotators agreed on their labels for the first two passages (A and B); the third passage (C) received a mix of the labels INFORMATION and DESCRIPTION, plus REPORT.

A. DESCRIPTION
The red house was surrounded by a large garden with lots of flowerbeds, fruit bushes, fruit trees of different kinds, a spacious lawn with a glider and a little gazebo that Granddad had built for Granny...

B. INFORMATION
The Group has three control functions, which are independent from the business operations: Internal Audit, Compliance and Risk control.

C. INFORMATION/REPORT
According to Chris Wille, the Rainforest Alliance's Chief of Sustainable Agriculture, technological advances and a more favorable market should facilitate a steady evolution toward ever better conditions on certified farms.

The intuitive descriptions we gave to the annotators intentionally avoided mentioning specific linguistic characteristics of the modes, and this may be one reason some distinctions were difficult to make. INFORMATION was frequently mentioned by annotators as the most confusing category and the most difficult to differentiate from the others. The choice to use paragraph boundaries instead of true DM boundaries also influenced the annotation process, as inspection of the most-disagreed-upon passages shows that many paragraphs in fact display a mix of DMs. Finally, several annotators seemed to have trouble making the distinction between labeling the DMs of individual passages rather than the genre of texts.

Cross-linguistic comparison. The next question to be addressed is to what extent DMs correspond across parallel aligned paragraphs of text for the language pair English-German. Given the differences between annotators, of course, these results can only be seen as suggestive. Table 2 shows the confusion matrix for DM annotations across languages, aggregated across all texts. Interestingly, the same mode pairs that were reported as being difficult to distinguish by individual annotators show the highest degree of confusability when we compare annotations across languages (light-grey shaded boxes in Table 2).

Additional annotations and more systematic investigation are needed in order to determine whether these patterns reflect preferences of individual annotators or rather differences in how the two languages realize discourse modes.

4 Situation entities: annotation scheme

The second focus of this study is the question of how *situation entity (SE) types*, as defined by Smith (2003), differ cross-linguistically, focusing on two closely-related languages, English and German. During annotation, we follow the existing scheme for English data (Friedrich and Palmer, 2014), and our own adapted scheme for German data (Section 4.1). Here, we give a brief description of the SE types relevant to this study (see the cited paper for more details).

- STATE: clauses introducing properties (*Mary is tall*); modalized clauses (*Mary can swim*); perfect tense (*Mary has submitted the paper)*
- EVENT: dynamic events, particular things that happened (*Mary ate a cupcake)*
- GENERALIZING SENTENCE: clauses reporting on regularities related to particular individuals (*Mary cycles to work*)
- GENERIC SENTENCE: clauses making statements about kinds (*Monkeys like bananas*)
- QUESTION: *Do you really need an example?*
- IMPERATIVE: *Hand me the pen!*

In the remainder of this section, we describe (a) our adaptation of this scheme to German, and (b) an experiment studying the interpretation of the German perfect as having stative or event readings, as this is a crucial difference when determining the SE types for English or German text.

4.1 Annotation scheme for German

For adapting the scheme, we first asked several annotators, German native speakers, to apply the existing English scheme to German data and report on problems they found. In addition, disagreements between the annotators were carefully analyzed. The scheme was then adapted to German in order to account for the identified differences between the two languages, outlined below.

Perfect tense. Possibly the most striking relevant difference between English and German is the interpretation of perfect tense. While in English, all clauses in perfect tense are interpreted as stative (Katz, 2003), the German perfect can have a stative or an event reading (fulfilling a function similar to the English simple past), or even be underspecified depending on the context. In English, SE annotators are instructed to label all clauses in perfect tense as STATEs; in German, annotators can label them as EVENT or STATE, depending on what they find to be appropriate. We introduce a new label EVENT-PERF-STATE for underspecified cases. In Section 4.2, we conduct an experiment studying in detail the interpretation of the German perfect with regard to stative/event readings; the findings there validate our choice to allow variable SE annotations for the German perfect.

Genericity of main referent. A clause's main referent is the entity 'the clause is about.' GENERIC SENTENCES have *generic* main referents, which are defined as references to kinds, and all other SE types have *non-generic* main referents. In English clauses, the main referent usually coincides with the grammatical subject, but this simple heuristic does not always apply for German. We identify the following cases where it can be difficult in German to select the main referent.

Examples (1) and (2) illustrate usages of the *impersonal passive*, which can be formed in German (unlike English) for intransitive verbs. The pronoun *es* is a grammatical placeholder, and annotators have to infer the main referent from the clause's discourse context. In (1), the first clause introduces a particular situation, and we can infer in the second clause that some particular group of people is talking again. In (2), again context determines the habitual/generic reading of the second clause.

(1) *(a) Jetzt ist Pause, (non-generic,* STATE*)*
(b) es wird wieder geredet. (non-generic, EVENT*)*
There's a break now, people are talking again.

(2) *(a) Früher gab es keine Nähmaschinen, (generic,* GENERIC SENTENCE*)*
(b) heute wird anders genäht. (generic, GENERIC SENTENCE*)*
In the past, there were no sewing machines, today one sews differently / sewing is done differently.

In addition, there is a group of impersonal perception verbs which are usually expressed with stative verbs, and require an argument either in dative, as in (3a), or accusative, as in (3b). In both cases, the argument in dative or accusative is considered to be the main referent of the clause.

(3) *(a) Es graut mir vor morgen. (non-generic,* STATE*) I dread tomorrow.*
(b) Mich friert es oft. (non-generic, GENERALIZING SENTENCE*) I often freeze.*

Statal passive. The statal passive (4a), in contrast to the processual passive (4b), focuses on the result or the "state" reached after a process, and are marked as STATEs.

(4) *(a) Die Tür ist geöffnet. (*STATE*)*
*(b) Die Tür wurde geöffnet. (*EVENT*)*
(a) The door is open. / (b) The door was opened.

Modal constructions. Modalized clauses describe, among others, possibilities, necessities or conditions rather than actual events, and are therefore marked as STATEs.[6] In German, two common constructions indicating necessity are *haben/sein + zu + infinitive*; these are similar to the English *have to + infinitive / is to be + past participle*. The *sich lassen* construction (5) indicates possibility.

(5) *Dieser Konflikt* ***lässt sich*** *ohne Gewalt* ***lösen****. (*STATE*)*
This conflict ***can be*** *solved without violence.*

[6]The coercions described here and in the following two paragraphs (subjunctive and *damit*) do not apply to GENERALIZING or GENERIC SENTENCES.

Subjunctive mood. The German *Konjunktiv* expresses doubt, possibility, speculations or conditionality. The verb construction *wir gehen in Urlaub* in (6) is dynamic, but the subjunctive mood coerces the SE type to be STATE.

(6) *(a)* ***Hätten*** *wir das Geld, (*STATE*)*
(b) ***gingen*** *wir morgen schon in Urlaub. (*STATE*)*
If we had the money, we ***would*** *go on holiday tomorrow.*

Final clauses with "damit". Final clauses (7) describe a purpose, an intention or a goal rather than an actual event, and are coerced to STATEs.

(7) *Erinnere mich nochmal, (*IMPERATIVE*)*
*damit ich pünktlich komme. (*STATE*)*
Remind me again so I will be on time.

Interim summary. We have now described the major differences identified when applying the English SE annotation scheme to German data. How clauses of each SE type are expressed is clearly language-dependent. However, our main finding is that the SE categories *are* applicable to German, and that the SE level of discourse analysis is cross-linguistically applicable. In the following section, we drill down on the most striking difference, the annotation of clauses in perfect tense.

4.2 Experiment on the interpretation of the German perfect by many annotators

German clauses in perfect tense may have either a temporal reading (past event, as in (8a)) or an aspectual reading indicating completedness of an event, as in (8b) (Klein, 2000).

(8) *(a) Gestern sind wir ins Kino gegangen.*
*(*EVENT*) Yesterday we went to the movies.*
*(b) Ich habe schon gegessen. (*STATE*)*
I have eaten.

The above examples clearly emphasize either the event or its result, but in some sentences, such as (9), it is hard to say which is more important; the construction is underspecified. For such cases, we introduce the label EVENT-PERF-STATE.

(9) *Sie haben mir den Job gegeben.*
*(*EVENT-PERF-STATE*)*
They gave me the job. / They have given me the job.

The focus of the experiment described in this section is to investigate to what extent German native speakers are able to agree on the relative salience of the state/event information. We conduct a large-scale experiment involving a large number of participants. To the best of our knowledge, interpretation of the perfect has not been investigated in this way for German before.

Experiment. The experimental data are 73 German sentences collected from several multilingual web sites. Two authors of this paper collaborated to provide reference labels for the sentences, marking 24 as STATE, 24 as EVENT and 25 as EVENT-PERF-STATE. We ask participants to give a rating for whether they think the state or the event matters more for a target word in a sentence (sentences are presented in their context, usually a very short paragraph). The rating scale is 1-5, where 1 means that only the event is important, and 5 means that only the state matters.

We recruit voluntary annotators via mailing lists of computational linguistics students at several German universities. We randomize the presentation of experimental items, ensuring that each annotation batch contains 1/3 STATE items, 1/3 EVENT items and 1/3 EVENT-PERFECT-STATE items. Each annotator is also shown four 'sample' items, two of which are clearly STATEs and two of which are clearly EVENTs. A total of 2,347 annotations were made by 102 German native speakers. To control for whether the participants read the short instructions carefully, we additionally exclude the data of participants who did not mark the two STATE samples with a score between 1-3, or the two EVENT samples as 3-5 on the scale. This reduced the data set to 1,611 annotations by 63 people. Each annotator marked 18 or more sentences (average: 25), and each sentence was annotated 13 or more times (average: 18).

Results. The averaged scores for each item can be seen in Figure 1. Towards either end of the scale, standard deviation is low, validating our hypothesis that some cases clearly have a preferred interpretation. For underspecified cases (i.e. those with means around 3), standard deviation is also high: many annotators only see one reading.

Most of the reference labels match the mean of the scores given by the annotators. However, there are some noticeable outliers. The EVENT item seen around the 70 mark on the x-axis is the

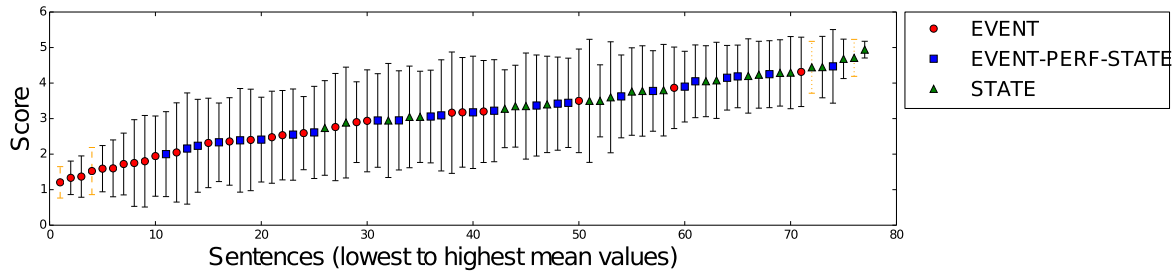

Figure 1: Perception of German perfect: mean and standard deviations for scores given to sentences. Semantics of scores: 1=EVENT, 3=EVENT-PERF-STATE, 5=STATE. Orange dotted lines: sample items.

sentence *Warum haben wir eigentlich geheiratet? (Why did we get married?)* – here, the state of being married is apparently quite prominent to our participants. The EVENT outlier seen around the 60 mark is *welches wir oben beschrieben haben (which we (have) described above)*, which probably should have been marked as EVENT-PERF-STATE.

Related work. A corpus study by de Swart (2007) analyzes the usage of the perfect in translations of the French novel *L'étranger* by Albert Camus. They show that the present perfect can be used to tell a story in French or German, but not in English or Dutch. Nishiyama and Koenig (2006) assess the role that the English perfect plays in discourse by examining the interpretations of 605 present perfect examples. Scheifele (2014) uses a picture-sentence-verification task to study the activation of the resultant state of sentences in the German perfect.

5 Situation entities: corpus study

With the annotation scheme established, we now compare the SE annotations on the parallel corpus.

5.1 Agreement

Each segment of the corpus data described in Section 2 is separately marked by three different annotators. Most annotations were done by paid, trained annotators, with some labels provided by one of the authors. Annotators were given the written manuals and trained on a few documents not included in the corpus. We create a gold standard via majority voting. The Smultron part of the data was labeled by a different combination of annotators than the rest of the documents. As Table 3 shows, substantial agreement was achieved. The categories (SE types) apply equally for both languages, but the mapping from linguistic structures to these types is language-dependent. The agreement numbers show that the two sets of guidelines work equally well for German and English.

corpus section	English	German
Smultron	0.63	0.62
other	0.61	0.67

Table 3: Agreement for SE type labels, Fleiss' κ.

5.2 Cross-lingual comparison of SE types

In this section, we move on to the cross-lingual comparison of the SE types of parallel texts. Our main questions are: do SE types in the texts of one language usually correspond to translated segments of the same SE type; and what are the cases in which aligned segments have different SE types? We use the subset of segments which have an aligned counterpart in the respective other language for this analysis.

As the confusion matrix of SE type labels in Table 4 shows, in most cases, the aligned segments receive the same SE type labels. This level of linguistic discourse analysis holds across languages and can potentially be relevant for improving or evaluating machine translation or translations of language learners: mismatches could be indicators for bad translations. However, mismatches can also occur for *good* translations in certain circumstances. In the following, we present a qualitative analysis of the non-matching cases with regard to whether they represent errors in annotation or patterns of SE type shift across languages.

Table 5 shows the counts of various mismatch types we identified for the aligned segments whose SE type labels differ. We found about 40% of mismatches to be results of disagreements, as they would occur in a monolingual setting as well.

		German							
		STATE	EVENT	EVT-PERF-ST	GENERAL.	GENERIC	IMP.	QUEST.	–
English	STATE	642	85	27	14	47	0	4	34
	EVENT	40	304	14	10	5	1	0	9
	GENERALIZING	9	5	0	38	49	1	0	6
	GENERIC	33	0	0	1	143	0	0	3
	IMPERATIVE	2	1	0	0	0	9	0	2
	QUESTION	2	0	0	0	1	0	62	5
	–	57	32	2	8	41	0	4	37

Table 4: Confusion matrix of SE type labels for parallel English-German text.

mismatch type	**#**
systematic disagreements	259
- involving German perfect	62
- lexical choice	79
- grammatical structure	5
- segmentation	113
language-independent disagreements	184
- genericity	125
- lexical aspectual class	17
- other	42

Table 5: Reasons identified for SE type mismatch.

For example, we find mismatches between judging the main referent of a segment as generic or non-generic, which has been found to be a difficult decision before (Friedrich et al., 2015). Most of these cases seem to be language-independent, however, there are cases where a certain form of the noun phrase primes a particular reading. The GENERIC SENTENCE *Terrorists may also benefit* has been translated as *Auch die Terroristen könnten profitieren* (STATE), in which the noun phrase *die Terroristen* primes the non-generic reading in this context. Further cross-linguistic study of the expression of generic noun phrases is needed.

About 60% of the disagreements were identified as resulting from cross-lingual differences. As expected, the German perfect causes confusion between STATEs and EVENTs. Additional confusion between these two types results from lexical choice, presenting the same matter of affairs as either a STATE or an EVENT, as in *She was startled* (STATE) vs. *Sie fuhr zusammen* (EVENT). Similarly, the lexical aspectual class of the English verb *support* can be interpreted as stative or dynamic, but the German translation *fördern* has a stronger preference for a dynamic interpretation.

Some clauses are marked as a segmentation error in one language, meaning the clause does not contain a full verb constellation. This occurs for example if a clause contains only an infinitive. If the other language did not use an infinitive construction, the segment receives a label.

In addition, requests can be formulated in different ways and lead to mismatches, as the following example shows: *Take a look at...* (IMPERATIVE) vs. *Hier können Sie ... sehen* (STATE).

This paper presents a small pilot study, but it shows clearly that some SE type shifts are systematic for this closely related language pair. As future work, we suggest investigating whether these SE type shifts can be predicted with an automatic classifier. This in turn could be a valuable resource for translation studies or for improving or evaluating machine translation.

6 Discourse modes and situation entities

We have studied the cross-linguistic correspondences of SE types and DMs above. The final step in the study brings these two levels of analysis together by looking at the distributions of SE types for paragraphs of different DM types. According to Smith, SE distributions should be one of the distinguishing features between text passages of different DM types: NARRATIVE and DESCRIPTION passages contain large numbers of EVENTs and STATEs; REPORT passages contain these two types plus GENERALIZING and GENERIC SENTENCEs; INFORMATION and ARGUMENT/COMMENTARY should contain higher proportions of the latter two SE types.

Annotators were not told which SE types to expect in DMs; DM annotation was done purely intuitively. Note that the SE annotations were created via majority voting using established annotation schemes, and are thus quite reliable, but the DM labels are the results of the pilot study on DM annotation as described in Section 3. Table 6 shows the percentage of clauses labeled with a given SE type for each DM. We exclude clauses that the SE annotators marked as segmentation errors but include those which received no SE label

DM	# clauses	% STATE	% EVENT	% EV-PRF-ST	% GNRL.	% GENERIC	% IMP.	% QST.	% –
NARR.	288 / 341	57 / 53	25 / 29	0 / 1	2 / 6	8 / 4	1 / 3	0 / 1	6 / 4
REPORT	503 / 220	59 / 54	26 / 29	0 / 2	5 / 7	5 / 5	1 / 1	0 / 0	4 / 2
INF.	613 / 726	58 / 46	14 / 25	0 / 1	5 / 20	15 / 4	1 / 0	1 / 0	6 / 3
DESCR.	280 / 341	61 / 46	21 / 23	0 / 1	4 / 18	5 / 6	2 / 2	1 / 0	6 / 4
ARG/COM	552 / 553	57 / 46	19 / 20	0 / 2	12 / 24	7 / 4	1 / 1	1 / 0	3 / 3
OTHER	19 / 101	90 / 48	11 / 19	0 / 7	0 / 16	0 / 4	0 / 2	0 / 1	0 / 4
NONE	70 / 72	36 / 38	23 / 29	0 / 6	27 / 13	6 / 5	3 / 4	3 / 0	3 / 6

Table 6: Distribution of SE type labels per DM, as **percentage (%)** of all clauses per DM: *(En / De)*

due to annotator disagreements (marked as –).

Discussion. The distribution of SE types largely matches the predictions of Smith (2003). In all DMs, the predominant SE type is STATE.[7] The reason is possibly that STATE marks several different types of coerced cases (e.g., perfect, negation, modals). In future work, we are planning to investigate the different types of STATEs per mode. There is already a clear path for this investigation: for each clause, we also annotated features such as the type of main referent, the lexical aspectual class of the main verb, and habituality (as described in the original annotation scheme by Friedrich and Palmer (2014)). These features will allow us to quickly sub-type clauses labeled STATE. EVENT-PERF-STATE of course appears only in the German data.

The most interesting differences show up in the distributions of the SE types which convey general rather than specific information: both GENERALIZING SENTENCE and GENERIC clauses figure more prominently in the modes of INFORMATION, DESCRIPTION, and ARGUMENT/COMMENTARY than they do in NARRATIVE or REPORT.

It should also be noted that the distributions shown here could to some extent be affected by problems with automatically aligning clauses across languages. The non-Smultron portion of the corpus is manually aligned, and there we retain from roughly 80-90% of the annotated clauses. The Smultron data is automatically aligned, and there we drop to below 60% of the clauses.

7 Conclusion and future work

The present corpus study shows that discourse analysis at the level of DMs and semantic analysis at the level of SEs are quite robust across the two closely related languages German and English. Both of these phenomena have been investigated from a theoretical perspective for other languages (Smith, 2003), with a small empirical study for Mandarin (Smith and Erbaugh, 2001), and further empirical analysis of additional languages is certainly warranted.

The DM annotation pilot study confirmed the expectation that paragraph boundaries as signaled by white space in the original documents do not correspond cleanly to actual DM borders, and these mixed paragraphs were especially difficult for annotators to label. Another question for future work is whether to allow one passage to have a mixture of DMs (for example, sometimes NARRATIVE passages have many background INFORMATION sentences), or whether additional DMs should be introduced.

Finally, as future work, we plan to create computational models of SEs and DMs, and exploit their relationship as empirically ascertained in Section 6. These computational models could then in turn be used to improve NLP applications as mentioned in the introduction.

Acknowledgments

First and foremost, thanks to the fleet of volunteer annotators who gave their time for either the discourse modes pilot annotation study or the German perfect study, as well as our excellent situation entity annotators. We also thank Alessandra Zarcone for invaluable help with experimental design, and the anonymous reviewers for their feedback. This work is funded in large part by the Cluster of Excellence *Multimodal Computing and Interaction* and in smaller part through the Leibniz ScienceCampus *Empirical Linguistics and Computational Language Modeling*. The second author is supported by an IBM PhD Fellowship.

[7] Although the proportion of STATEs appears to be unusually high for English paragraphs with the DM label OTHER, investigation of this data revealed no particular patterns. Instead, this is an anomaly due to the very small sample size.

References

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press, Cambridge.

Henriëtte de Swart. 2007. A cross-linguistic discourse analysis of the perfect. *Journal of pragmatics*, 39(12):2273–2307.

Annemarie Friedrich and Alexis Palmer. 2014. Situation entity annotation. In *Proceedings of the 8th Linguistic Annotation Workshop*.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*.

Zahurul Islam and Alexander Mehler. 2012. Customization of the Europarl Corpus for Translation Studies. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Graham Katz. 2003. On the stativity of the English perfect. *Perfect explorations*, pages 205–234.

Dan Klein and Christopher D Manning. 2002. Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems*, pages 3–10.

Wolfgang Klein. 2000. An analysis of the German Perfekt. *Language*, pages 358–382.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of EMNLP*.

Atsuko Nishiyama and Jean-Pierre Koenig. 2006. The perfect in context: A corpus study. *University of Pennsylvania Working Papers in Linguistics*, 12(1):22.

Alexis Palmer and Annemarie Friedrich. 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*.

Rebecca J. Passonneau, Nancy Ide, Songqiao Su, and Jesse Stuart. 2014. Biber Redux: Reconsidering Dimensions of Variation in American English. In *Proceedings of COLING*.

Marina Santini. 2006. Towards a zero-to-multi-genre classification scheme. In *Proceedings of ATALA*.

Edith Scheifele. 2014. The German Perfekt: Activation of the Resultant State? In *Linguistic Evidence 2014, Empirical, Theoretical, and Computational Perspectives*, Tübingen.

Carlota S. Smith and Mary S. Erbaugh. 2001. Temporal Information in Sentences of Mandarin. In Xu Liejiong and Shao Jingmin, editors, *New Views in Chinese Syntactic Research – International Symposium on Chinese Grammar for the New Millenium*.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*. Cambridge University Press.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of NAACL-HLT*. Association for Computational Linguistics.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Publications.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Martin Volk, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. 2010. SMULTRON (version 3.0) — The Stockholm MULtilingual parallel TReebank.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of ACL*.

Egon Werlich. 1975. *Typologie der Texte*. Quelle & Meyer, Heidelberg.