Automatic recognition of habituals: a three-way classification of clausal aspect

Annemarie Friedrich **Manfred Pinkal**
Department of Computational Linguistics
Saarland University, Saarbrücken, Germany
{afried,pinkal}@coli.uni-saarland.de

Abstract

This paper provides the first fully automatic approach for classifying clauses with respect to their aspectual properties as habitual, episodic or static. We bring together two strands of previous work, which address only the related tasks of the episodic-habitual and stative-dynamic distinctions, respectively. Our method combines different sources of information found to be useful for these tasks. We are the first to exhaustively classify *all* clauses of a text, achieving up to 80% accuracy (baseline 58%) for the three-way classification task, and up to 85% accuracy for related subtasks (baselines 50% and 60%), outperforming previous work. In addition, we provide a new large corpus of Wikipedia texts labeled according to our linguistically motivated guidelines.

1 Introduction

In order to understand the function of a clause within a discourse, we need to know the clause's aspectual properties. The distinction between *dynamic* and *stative* lexical aspect, as in examples (1a) and (1b) respectively, is fundamental (Vendler, 1957). Its automatic prediction has been addressed previously (Siegel and McKeown, 2000; Zarcone and Lenci, 2008; Friedrich and Palmer, 2014).

(1) (a) Bill drank a coffee after lunch. (**dynamic**)
(b) Bill likes coffee. (**stative**)

In this work, we focus on *habituality* as another fundamental aspectual property. While example (1a) is an *episodic* sentence, i.e., a sentence expressing information about a particular event, the same dynamic verb can be used to characterize the default behavior of an individual or of a kind in a certain type of situation (2).

(2) (a) Bill usually drinks coffee after lunch. (**habitual**)
(b) Italians drink coffee after lunch. (**habitual**)

The entailment properties of episodic and habitual (or *characterizing*) sentences differ substantially. Also, they have different functions in discourse. The episodic event expressed by (1a) is typically embedded in the temporal structure of a narration. The habitual sentence (2a) can be used, e.g., as an explanation (why Bill is still sitting at the table), or in a contrastive context (today, he ordered a grappa instead). *Generic* sentences with *kind-referring* subjects (2b) can also be habitual, generalizing at the same time over typical members of this kind and over situations in which they typically carry out some action.

Habitual sentences do not move narrative time, similar to stative clauses such as (1b). Carlson (2005) considers habituals to be aspectually stative. Since there are clear differences between ordinary statives such as (1b) and habituals, we apply a three-way distinction for *clausal aspect* in this work. We classify clauses as one of the three categories **habitual**, **episodic** and **static**.[1]

Through its impact on entailment properties and temporal discourse structure, the determination of clausal aspect is relevant to various natural language processing applications requiring text understanding, such as novelty detection (Soboroff and Harman, 2005), knowledge extraction from text (Van Durme, 2010) or question answering (Llorens et al., 2015). Using aspectual information has been shown to improve temporal relation identification (Costa and Branco, 2012).

Some languages (e.g., Czech or Swahili) have systematic morphological markers of habituality

[1] For clarity, we use the label *static* for the clausal aspect of non-episodic and non-habitual sentences. We reserve *stative*, which is more common in the literature, for the lexical aspectual class.

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2471–2481,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

(Dahl, 1985). In other languages, there are cues for habituality, such as the simple present in English, and the use of certain adverbials (Dahl, 1995). The automatic recognition of habitual sentences for the latter languages is non-trivial. The work in this paper targets the English language; we leave recognition of habituality in other languages to future work.

The only previous work on categorizing sentences as episodic or habitual we know of is by Mathew and Katz (2009). They do not attempt to categorize arbitrary sentences in 'free text', however, but work with a corpus of selected sentences and use gold standard parse information for their experiments. In particular, they consider clauses containing lexically dynamic verbs only.

In this work, we address the task of an exhaustive classification of all clauses of a text into the three aspectual classes **habitual**, **episodic**, and **static**. Static sentences include lexically stative clauses as well as negated or modalized clauses containing a dynamic main verb. A computational model for identifying episodic and habitual clauses clearly needs to address this third class as well if it is to be applied in a realistic setting. Linguistically, the determination of clausal aspect depends on the recognition of the verb's lexical aspectual class (stative or dynamic), and on the recognition of any aspectual markers or transformations, such as use of the perfect tense, negations or modals. Our work builds on results for the related subtasks (Mathew and Katz, 2009; Siegel and McKeown, 2000; Friedrich and Palmer, 2014), using both context-based and verb-type based information.

Our major contributions are: (i) We create a corpus of 102 Wikipedia texts whose about 10,000 clauses are annotated as episodic, static or habitual with substantial agreement. This corpus allows for studying the range of linguistic phenomena related to the clause types as defined above (compared to previous work which uses only a small set of verbs and sentences), and provides a basis for future research. (ii) We provide the first fully automatic approach for this classification task, combining two classification tasks (lexical aspectual class and habituality) that have been treated separately in previous work. For an exhaustive classification of clauses of free text, these two classification tasks need to be addressed jointly. We show two different feature sets (verb-type based features and context-based features) to have different impact on the two subtasks, and to be complementary for our full three-way task.

We reach accuracies of nearly 85% for the two subtasks of identifying static clauses and distinguishing episodic and habitual clauses (majority class baselines are 60% and 50% respectively). A joint model for the three-way classification task reaches an accuracy of 80% (baseline 60%). In addition, we show that the verb-type based linguistic indicator features generalize well across verb types on our tasks: for verbs unseen in the training data, accuracies drop only by 2-5%.

2 Theoretical background and annotation guidelines

In this section, we give an overview of the semantic theory related to habituals, at the same time introducing our annotation guidelines for marking clauses as **habitual**, **episodic** or **static**.

2.1 Habituality

Habitual sentences express regularities in terms of generalizations over events and activities. In semantic theory, habituals are formally represented using a quantifier-like operator GEN (Krifka et al., 1995):

(3) GEN[s](s **is an after-dinner situation** & Bill **is involved in** s; Bill **drinks a coffee in** s)

In the semi-formal representation (3) of sentence (2a) above, GEN binds a variable s ranging over situations, the first argument restricts the situation type, and the second argument provides the activity that is typically carried out by the protagonist in the respective situations. The GEN operator is similar to the universal quantifier of predicate logic. However, habitual sentences tolerate exceptions: (2a) is true even if Bill does not drink a coffee after every lunch. Also note that habituals are not restricted to what one would consider a matter of habit; they can also have inanimate subjects, as illustrated by (4).

(4) Glass breaks easily. (**habitual**)

2.2 Clausal and lexical aspectual class

Clausal aspect is dependent on the lexical aspectual class (stative or dynamic), but the two levels are essentially different. Dynamic verbs express events or activities (e.g., *kill, fix, walk, forget*), while stative verbs express states (e.g., *be,* *like, know, own*). The *fundamental aspectual class* (Siegel and McKeown, 2000) of a verb in context describes whether it is used in a stative or dynamic sense before any aspectual markers or transformations (such as use of the past/present perfect or modals) have been applied. It is a function of the main verb and a select group of complements (it may differ per verb which ones are important). For example, the fundamental lexical aspectual class of the verb *make* with the subject *Mary* and the object *cake* in (5) is dynamic. English clauses in past or present perfect such as (5) are static, as they focus on the post-state of an event rather than the event itself (Katz, 2003).

(5) Mary has made a cake. (**static**)

Habituals with verbs of dynamic aspectual class are by far more frequent in our corpus,[2] but there are also instances of stative verbs used in a habitual way, as for example (6).

(6) Sloths sometimes sit on top of branches.
(**habitual**, *stative* lexical aspectual class)

2.3 Modality and negation

Modalized (7) and negated sentences (8) tend to be static: they do not express information about a particular event, but refer to actual or possible states of the world.

(7) Mary can swim. (**static**)

(8) Mary didn't go swimming yesterday. (**static**)

The above definitions of habituality and stativity are generally agreed upon in literature. However, the interaction of these phenomena is by no means trivial (Hacquard, 2009), and required making some decisions during the design of our annotation guidelines. Here, we explain these decisions, which are all motivated by a clause's entailment properties.

One difficult issue is how to interpret and mark negated sentences such as (9a) whose positive version (9b) is habitual.

(9) (a) John does not smoke. (**habitual**)
(b) John smokes. (**habitual**)

Sentence (9a) can be considered either **static** because of the negation (*It is not the case that John smokes*), or as **habitual** because it characterizes John's behavior (*In any relevant situation, John does not smoke*). Both decisions are possible (Garrett, 1998), we decide for the latter possibility. This decision is supported by the observation that (9a) is similar in its entailment properties to (10), which due to the frequency adverbial *never* clearly generalizes over relevant situations (though note that this is not a linguistic test).

(10) John never smokes. (**habitual**)

Likewise, we mark modalized sentences as habitual if they have a strong implicature that an event has actually happened regularly (Hacquard, 2009), as in (11). In contrast, (7) is static as it does not imply that Mary actually swims regularly.

(11) I had to eat an apple every day. (**habitual**)

The above example shows that modality and habituality are interweaved and sometimes hard to identify. Nevertheless, we reach substantial agreement in the annotation of our corpus (see Section 4.2).

Finally, some habituals have a *dispositional* reading, indicating ability/capability (Menéndez-Benito, 2012). Example (12) can be paraphrased by (13), as it does not indicate that the car is actually driven this fast regularly, it only states its maximum speed.

(12) This car goes 200 kph.

(13) This car can go 200 kph.

3 Related work

The task of predicting *fundamental aspectual class* aims to determine whether the verb is used in a stative or dynamic sense. This task predicts the aspectual class of a verb in context before any aspectual markers or transformations (such as use of the perfect or modals) have been applied. Siegel and McKeown (2000) propose the use of linguistic indicators (explained in Section 5.2); Friedrich and Palmer (2014) show the importance of using context-based features in addition. Zarcone and Lenci (2008) classify occurrences of 28 Italian verbs according to Vendlers classes state, process, accomplishment and achievement.

Mathew and Katz (2009) address the problem of 'supervised categorization for habitual versus

[2] The distribution of lexical aspectual class of verbs is generally skewed towards dynamic (Friedrich and Palmer, 2014).

episodic sentences' . The authors randomly select 1052 sentences for 57 verbs from the Penn TreeBank (Marcus et al., 1993) and manually mark them with regard to whether they are **habitual** or **episodic**. They focus on verbs that are lexically dynamic and discuss a variety of syntactic features, which they extract from gold standard parse trees. Their aim is to study the ability of syntactic features alone to identify habitual sentences.

Xue and Zhang (2014) annotate verbs with the four event types *habitual event*, *state*, *on-going event* and *episodic event* with the aim of improving tense prediction for Chinese. Recent related work (Williams, 2012; Williams and Katz, 2012) extracts typical durations (in term of actual time measures) for verb lemmas from Twitter. They distinguish episodic and habitual uses of the verbs, using the method of Mathew and Katz (2009).

4 Data

In this section, we describe the data sets used in our experiments.[3]

4.1 Penn TreeBank (M&K) data set

Mathew and Katz (2009) randomly select sentences for several verbs from the WSJ and Brown corpus sections of the Penn Treebank. They require the verb to be lexically dynamic. Sentences are marked as **habitual** or **episodic**, further details on the annotation guidelines are not specified. Their data set contains 2743 annotated sentences for 239 distinct verb types. Mathew and Katz remove verb types with highly skewed distributions of labels, but their filtered data set is not available. We follow their filtering approach, but we could not replicate their filtering step. Our final data set contains 1230 sentences for 54 distinct verb types. Mathew and Katz (2009) state that their data set comprises 1052 examples for 57 verb stems. We aimed at producing a similar distribution of labels: our data set contains 73.3% episodic cases, M&K's version has 73.1%.

4.2 Wikipedia corpus

We select 102 texts from a variety of domains from Wikipedia, as we expect an encyclopedia to contain many habitual sentences. We use the discourse parser SPADE (Soricut and Marcu, 2003) to automatically segment the first 70 sentences of each article into clauses. A clause is approximately defined as a finite verb phrase. Each clause is then labeled as **static**, **episodic** or **habitual**. Details on our annotation scheme have been given in Section 2. Annotators are allowed to skip nonfinite clauses (e.g., headlines only containing a noun phrase). This happened in about 14% of all pre-segmented clauses. The final Wikipedia data consists of 10355 labeled clauses. Table 1 gives statistics for the distribution of labels.

Label	#	%
static	6184	59.7
episodic	2114	20.4
habitual	2057	19.9
total	10355	-

Table 1: Wikipedia data, distribution of labels for clausal aspect.

The data set was labeled by three paid annotators, all students of computational linguistics. Annotators were given a written manual and a short training on documents not included in the corpus. Agreement on the Wikipedia data is 0.61 in terms of Fleiss' κ, which indicates substantial agreement (Landis and Koch, 1977). The gold standard that we use in our experiments is constructed via majority voting. The gold standard contains the cases where at least two annotators agreed on the label. We found only 86 cases where all annotators disagree, and manual inspection shows that most of these cases are related to disagreements on the lexical aspectual class that coincide with an attention slip by one of the annotators.

5 Method

In this section, we describe our computational models for determining clausal aspect.

5.1 CONTEXT-BASED features

Table 2 shows the syntactic-semantic features, which we call CONTEXT-BASED as they are extracted from the context of each verb occurrence that we classify. This feature set comprises the features proposed by Mathew and Katz (2009) and the ones proposed by Friedrich and Palmer (2014). In addition, we use the features *modal* and *negated*. We extract these features from syntactic dependency parses created using the Stanford parser (Klein and Manning, 2002). Tense and voice are extracted following the rules pro-

[3] All data sets are freely available from `www.coli.uni-saarland.de/projects/sitent`. We thank Thomas A. Mathew and Graham Katz for allowing us to publish their data set.

Feature		Values
verb	tense*†	past, present, infinitive
	pos†	VB, VBG, VBN, ...
	voice†	active, passive
aspect	progressive*†	true, false
	perfect*†	true, false
subject	bare plural*	true, false
	definite*	true, false
	indefinite*	true, false
object	absent*	true, false
	bare plural*	true, false
	definite*	true, false
	indefinite*	true, false
grammatical dependents†		WordNet lexname/POS
sentence	modal	*would, can,...*
	negated	true, false
	conditionals*	presence of clause starting with if/when/ whenever
	temporal modifiers*	specific, quantificational, including *used to* and *would* (where no if)
	prepositions*	at / in / on (3 features, true/false)

Table 2: CONTEXT-BASED features. Used by: *Mathew and Katz (2009), †Friedrich and Palmer (2014).

vided by Loaiciga et al. (2014). The values of the grammatical dependents' features are the WordNet (Miller, 1995) lexical filename of the dependent's lemma, or, if not available, the dependent's part-of-speech tag. Quantificational adverbs are temporal modifiers such as *always*, *occasionally* or *weekly*.[4] Specific temporal adverbs are, according to a heuristic proposed by Mathew (2009), phrase children marked with the part-of-speech tag TMP and whose child is a prepositional phrase. Noun phrases with one of the determiners *the, this, that, these, those, each, every, all*, as well as possessives, pronouns, proper names and quantified phrases are definite. NPs with determiners *a, an, many, most, some*, and cases of modifying adjectives without determiners (e.g., *few*) or cardinal numbers (part-of-speech tag CD) are indefinite. Mathew (2009) describes their features in detail.

5.2 TYPE-BASED features

This feature set consists of the verb-type based linguistic indicator features of Siegel and McKeown (2000). The computation of these features is based on a large parsed, but otherwise unannotated background corpus. For each verb type (i.e., lemma), these features count how often the verb occurs with each of the linguistic indicators as listed in Table 3. Except for the frequency feature, these values are normalized by the number of occurrences of the verb type. For example, if the verb type *win* occurs 1000 times in the parsed background corpus, of which 100 times with perfect aspect, the value of the linguistic indicator feature `perfect` is 0.1 for the verb type *win*. For any instance whose verb's lemma is *win*, 0.1 will be the value of the feature `perfect`, in other words, all instances of the same verb type receive the same TYPE-BASED feature values. Linguistic indicator features have recently been applied successfully on the related task of classifying the lexical aspectual class of verbs by Friedrich and Palmer (2014), who extract the linguistic indicators from an automatically parsed version of the AFE and XIE parts of Gigaword. We use their database of linguistic indicator values.[5]

Feature	Example	Feature	Example
frequency	-	continuous adverb	*continually* *endlessly*
present	*says*	evaluation adverb	*better* *horribly*
past	*said*	manner adverb	*furiously* *patiently*
future	*will say*	temporal adverb	*again* *finally*
perfect	*had won*	in-PP	*in an hour*
progressive	*is winning*	for-PP	*for an hour*
negated	*not/never*		
particle	*up/in/...*		
no subject	-		

Table 3: TYPE-BASED feature set (**linguistic indicators**) and examples for lexical items associated with each indicator, following Siegel and McKeown (2000).

5.3 Algorithm

In order to investigate in which circumstances the task of predicting a clause's label (**habitual**, **episodic** or **static**) can be addressed jointly, or whether a pipelined approach is better, we apply the following methods. Our JOINT model learns the decision boundaries for the three classes jointly, i.e., as a three-way classification task. In addition, we present a CASCADED model, which uses two models learned for the two different subtasks: (a) identifying static clauses and (b) distinguishing episodic and habitual clauses.

First, we train a model to distinguish the **static** class from the other two. In this learning step, we simply map all the clauses labeled as **episodic** and

[4] The complete list of quantificational adverbs used is given by Mathew (2009), page 36.

[5] `www.coli.uni-saarland.de/projects/sitent`

habitual to the class **non-static** and learn the decision boundary between the two classes **static** and **non-static**. Second, we train a model to distinguish the **episodic** from the **habitual** class. This model is trained on the subset of examples labeled with either of these two classes.

In the CASCADED model, first, the **static** vs. **non-static** model is applied. The CASCADED model labels all instances automatically labeled as **static** in this first step, and then applies the second model (**episodic** vs. **habitual**) on all remaining instances.

We train Random Forest classifiers (Breiman, 2001) using Weka (Hall et al., 2009) for each step and also for the JOINT model. Besides providing a robust performance, Random Forest classifiers can easily deal with both categorical and numeric features. This is relevant as our CONTEXT-BASED features are categorical while the TYPE-BASED features are numeric. In our experiments, we will compare the impact of the different feature sets on each subtask and on the JOINT model.

5.4 Baseline: Mathew and Katz (2009)

As a baseline, we also report results for the subset of our CONTEXT-BASED features used by Mathew and Katz (2009) and call this subset MK. Mathew and Katz (2009) find a J48 decision tree and a Naive Bayes classifier to work best. We replicate their results for the decision tree in Section 6.2.

6 Experiments and discussion

This section describes our experiments. First, we reproduce the experiments of Mathew and Katz (2009), who use manually created syntactic parses, in a purely automatic setting.

The data set and experiments of Mathew and Katz (2009) focus on the episodic-habitual distinction using a set of sentences selected for a small set of verbs, and their feature design focuses on syntactic properties of the clauses found in this annotated data set. In the further experiments, we turn to the Wikipedia data, which contains annotations for full texts. We expect the Wikipedia data to cover the range of habitual and episodic expressions more fully, and in addition, allows for studying the task of separating static sentences from the other two classes. As we will show, this latter task profits from including features relevant to the stative-dynamic distinction on the lexical level.

We first present experimental results for the two subtasks (described in Section 5.3). Our CASCADED model first identifies static clauses, and then classifies the remaining clauses as episodic or habitual. For reasons of readability, we first report on our experiments for the episodic-habitual distinction using both the M&K and Wikipedia data sets. Using the Wikipedia data, we then report on the results for the static vs. non-static distinction. Finally, we turn to the full task of the three-ways distinction.

6.1 Experimental setting

We report results for 10-fold cross validation (CV) with two different settings: In the RANDOM CV setting, we randomly distribute the instances over the folds, putting all instances of one document into the same fold. In the UNSEEN VERBS CV setting, we simulate the case of not having labeled training data for a particular verb type by putting all instances of one verb type into the same fold.

We compute the information retrieval statistics of precision (P), recall (R) and F1-measure per class, where F1 is the harmonic mean of P and R, $F1 = \frac{2*P*R}{P+R}$. Macro-average P is computed as the (unweighted) average of the P scores of the classes, and macro-average R is computed likewise. Macro-average F1 is the harmonic mean of macro-average P and macro-average R. We use McNemar's test with $p < 0.01$ to compute statistical signficance of differences in accuracies. In our tables, we indicate that two results differ significantly by marking them with the same symbols (we only show this when scores are close).

6.2 M&K data: episodic vs. habitual

We use Weka's 10-fold stratified cross validation and a J48 decision tree in the experiments reported in this section in order to replicate their experimental setting. Results are shown in Table 4. For the sake of completeness, we also show the results as presented in the original paper. F1-scores are computed from P and R as reported in the original paper. Note that their experiments are performed on a different subset of the data and so these numbers are not directly comparable to ours, but our subset has a very similar class distribution (see Section 4.1). Our accuracies based on automatic parses rather than gold standard parses are about 3% lower when using the original feature set (MK). We conclude that our results are in the expected range. Also, we do not find any significant improvements on this data set when using any other feature sets or combinations thereof (the table shows the results for our Context-based feature set); the M&K feature set designed for this corpus captures its variation well.

We have used a J48 decision tree in this section for comparability with previous work. In all following sections, we present results using Random Forest classifiers as described in Section 5.3.

	F1-score			**Acc.**
System	epis.	habit.	*macro*	
majority class*	84.5	0.0	42.2	73.1
MK*	91.1	70.5	80.8	86.1
MK	89.6	63.5	76.5	83.8
Context	90.0	64.7	77.3	84.4

Table 4: Results for **episodic** vs. **habitual**, J48 decision tree, data from Mathew and Katz (2009). *Numbers from original paper.

6.3 Wikipedia: episodic vs. habitual

We study the classification task of distinguishing **episodic** and **habitual** sentences using the subset of the Wikipedia data having one of these two labels (4171 instances). This task parallels the experiment of Mathew and Katz (2009) described above. We conduct two experiments, once using the Random CV setting and once using the Unseen verbs setting. Table 5 shows the results. The distribution of instances is nearly 50:50 in the gold standard (see Section 4, Table 1), and the majority classes in the respective training folds differ (this is the reason for the different baseline scores). For reasons of space we do not show the other scores here; macro-average F1-scores have (almost) the same values as accuracy, the F1-scores for episodic and habitual are similar to each other in each case.

Our findings are as follows: Type-based features outperform the majority class baseline, which means that some verbs have a preference for being used as either episodic or habitual. The Context-based features work remarkably well. If training data of the same verb type is available, adding the Type-based features or the lemma to the Context-based features results in improvements; this is not the case in the Unseen verbs setting. The latter setting shows that the additional contextual features (compared to the MK subset) are important: our corpus indeed covers a broader range of phenomena than the M&K data set.

Features	Random CV	Unseen verbs
majority class	42.1	46.3
lemma	65.4	46.3
Type	68.1	53.9
MK	82.3	‡81.4
Context	*†82.8	‡**83.8**
+ lemma	*84.3	
Context + Type	†**85.1**	83.1
+ lemma	84.0	

Table 5: **Wikipedia**: **Accuracy** of **episodic** vs **habitual**, 4171 instances, 10-fold cross validation, *†‡differences statistically significant.

	Random CV		Unseen verbs	
Features	**F1**	**Acc.**	**F1**	**Acc.**
majority class	37.4	59.7	37.4	59.7‡
MK	67.5	*69.5	59.2	62.7‡
Context	70.3	*71.7	62.8	64.9‡
+ lemma	81.9	†82.8		
Type	78.8	79.3	72.2	73.2‡
Context + Type	83.6	†84.1	**78.4**	**79.2**‡
+ lemma	**83.8**	**84.4**		

Table 6: **Wikipedia**: **static** vs **non-static**. All 10355 instances, 10-fold cross validation.*†‡ differences statistically significant.

6.4 Wikipedia: static vs. non-static

We evaluate the task of classifying **static** versus **non-static** clauses using all 10355 instances of the Wikipedia data set. Any instance labeled **episodic** or **habitual** receives the label **non-static** both in training and testing. Results of this task are shown in Table 6. For this subtask, the Context-based features are less informative than the Type-based features. Again, using lemma information approximates the use of type-based information, but this is not an option in the Unseen verbs setting. A combination of the Context-based and Type-based features achieves the best results. Friedrich and Palmer (2014) find that Type-based features generalize well across verb types when predicting the aspectual class of verbs in context, the same is true here. They achieve small improvements by adding context-based features. Predicting the lexical aspectual class of the clause's main verb is only part of our classification task, the **static** class includes not only lexically stative clauses but also clauses with lexically dynamic verbs that are stativized, e.g., modals, negation or perfect tense. Hence, as expected, in our task, adding the Context-based features results in a considerable performance improvement (5-7% absolute in accuracy).

	Random cross validation					Unseen verb types experiment				
	F1-score				Acc.	F1-score				Acc.
Features	stat.	epis.	habit.	*macro*		stat.	epis.	habit.	*macro*	
majority class baseline	74.8	0	0	24.9	59.7	74.8	0.0	0.0	24.9	‡59.7
Joint: MK	76.6	65.4	26.1	57.5	*67.0	76.3	41.7	0.8	49.0	‡63.8
Joint: Context	77.5	65.8	36.4	60.5	*68.4	74.7	57.1	12.0	51.7	⋆63.9
+ lemma	85.5	75.0	51.6	71.8	†78.0					
Joint: Type	81.9	52.7	49.7	61.5	69.9	74.9	4.2	2.8	40.7	⋆60.0
Joint: Context + Type	86.1	75.8	58.8	73.8	†79.0	81.2	69.5	31.3	63.6	**72.1
+ lemma	86.8	75.0	59.9	74.2	79.6					
Cascaded	**86.9**	**76.1**	**62.2**	**75.1**	**79.9**	**82.6**	**72.0**	**50.2**	**68.4**	****74.3**

Table 7: **Wikipedia**: **static** vs. **episodic** vs. **habitual**. 10355 instances, 10-fold cross validation. The Cascaded model uses the best models from Table 6 and Table 5. *† ‡ ⋆** differences statistically significant.

It is worth noting that even for verbs not seen in the training data, high accuracies and F1-scores of almost 80% can be reached.

6.5 Wikipedia: combined task

In this section, we describe our experiments for the three-way classification task of **static**, **episodic** and **habitual** clauses, as in a realistic classification setting, a clause may belong to either of these *three* classes. We investigate whether a pipelined Cascaded approach is better, or whether the Joint model profits from learning the decision boundaries between all three classes jointly. The results for this task are presented in Table 7. Both the Context-based and the Type-based features when used alone improve over the majority class baseline by about 10% in accuracy in the Random CV setting, and only by about 4% in the Unseen verbs setting. In the latter setting, all feature sets when used alone are ineffective for identifying habituals. This indicates that the Context-based features only 'pick up' on some type-based information in the Random CV case. The best models for this Joint classification task use both the Context-based and the Type-based feature sets: F1-scores and accuracy increase remarkably. Again, in the Random CV setting, using the lemma results in a large performance gain, though using the Type-based features is beneficial, and, in the Unseen verbs setting, essential.

We apply the Cascaded model as described in Section 5.3, training and testing the models for the subtasks in each fold. In the Random CV setting, the accuracy of the Cascaded approach is not significantly better than the one of the Joint approach, though F1-scores for the less frequent **episodic** and **habitual** classes both increase. In the Unseen verbs setting, however, the difference is remarkable: macro-average F1-score increases by almost 5% (absolute) and accuracy increases by 2.2%. Most notably, the F1-score for the habitual class increases from 0.31 to 0.50 (due to an increase in recall). To conclude, the Cascaded approach is favorable as it works more robustly both for verb types seen or unseen in the training data.

6.6 Feature ablation

In the above sections, we have compared the two major feature groups of Context-based and Type-based features. In addition, we ablate each single feature from the best results for each experiment. For all classification tasks, we found features reflecting tense and grammatical aspect to be most important, both for the Context-based and Type-based features. In general, we observe that no single feature has a big impact on the results, accuracy drops only by at most 1-2%. This shows that our feature set is quite robust and some of the features (e.g., part-of-speech tag of the verb and tense) reflect partially redundant information. However, using only the best features results in a significant performance drop by several percentage points in the various settings, which means that though single features may not have a large impact, overall, the models for this classification task profit from including many diverse features.

For the **episodic-habitual** distinction in the Unseen Verbs setting, the definiteness of the object was an important Context-based feature. In the **static vs. non-static** task, the subject also plays an important role, as well as the Type-based feature for *continuous adverbs*. In the Unseen verbs setting, many Type-based features are important, including those indicating how often the verb type occurs with *adverbs of manner*, *negation* and *in-PPs* in the background corpus. For the **combined** three-way task, we found the main verb's lemma and the direct object to have most impact. Of the TYPE-BASED features, the *for-PP*, *present* and *temporal adverbial* were most important. In the UNSEEN VERBS setting, many linguistic indicator features (among others *past*, *progressive*, *negation*) play a greater role, as well as information about the object, subject and tense.

7 Conclusion

In this paper, we have presented an approach for classifying the aspect of a clause as **habitual**, **episodic** or **static**. Clearly, when exhaustively classifying all clauses of a text, the **static** class cannot be ignored; we have shown that we can separate these instances from episodic and habitual instances, most of which are lexically dynamic, with high accuracy. Our model for distinguishing episodic and habitual sentences integrates a wide range of contextual information and outperforms previous work. Previous work has only addressed the classification of lexical aspectual class and the automatic distinction of episodic and habitual sentences. Our work is the first bringing together two strands of work relevant to classifying clausal aspect, and we have shown that sources of information relevant to these two underlying aspectual distinctions are relevant for our three-way classification task.

We have shown that for distinguishing static sentences from the other two, TYPE-BASED *and* CONTEXT-BASED information is needed; for distinguishing episodic and habitual clauses, CONTEXT-BASED features are most important. Our experimental results show that the three-way classification task is most effectively approached by combining both contextual and verb-type based information. Especially for verbs unseen in the training data, we found the CASCADED approach to work better. It is hard for the JOINT approach to identify habitual clauses; while in the CASCADED approach, performance for both steps is high and adds up.

We found the overall performance of this task to be about 80% accuracy, and 75% macro-average F1-score. These scores suggest that this method may be usable as a preprocessing step for further temporal processing.

8 Future work

Our models do not yet take discourse information into account. Consider example (14) by Mathew and Katz (2009): The second sentence is habitual, but the only indicator for this is sentence-external.

(14) John rarely ate fruit. He just ate oranges. (**habitual**)

In some preliminary experiments, we tried to leverage the discourse context of a clause for its classification by means of incorporating the gold standard label of the previous clause as a feature. This did not result in significant performance improvements. However, further experiments trying to incorporate discourse information are due, and, due to our new corpus of fully annotated texts, now possible.

Another related research direction is the classification of the different types of static clauses, e.g., the different senses of modality (Ruppenhofer and Rehbein, 2012). As mentioned before, a finer classification of the temporal structure of clauses is needed, among others identifying the lexical aspectual class as well as *viewpoint aspect* as *perfective* vs. *imperfective* (Smith, 1997).

Finally, the next steps in this line of research are to integrate the aspectual information attributed to clauses by our model into models of temporal discourse structure, which in turn are useful for information extraction and text understanding tasks in general. Costa and Branco (2012) are the first to show that aspectual information is relevant here; we hope to show in the future that temporal processing profits from integrating more fine-grained aspectual information.

Acknowledgments

We thank the anonymous reviewers, Alexis Palmer, Alessandra Zarcone and Andrea Horbach for their helpful comments related to this work, Graham Katz for providing us with their data set, and our annotators Christine Bocionek, Kleio-Isidora Mavridou and Melissa Peate Sørensen. This research was supported in part by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative (DFG), and the first author is supported by an IBM PhD Fellowship.

References

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Greg Carlson, 2005. *The Encyclopedia of Language and Linguistics*, chapter Generics, Habituals and Iteratives. Elsevier.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 266–275. Association for Computational Linguistics.

Östen Dahl. 1985. *Tense and aspect systems*. Oxford, Blackwell.

Östen Dahl. 1995. The marking of the generic/episodic distinction in tense-aspect systems. *The generic book*, pages 412–425.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore, USA*.

Andrew Garrett. 1998. On the origin of auxiliary do. *English Language and Linguistics*, 2(02):283–330.

Valentine Hacquard. 2009. On the interaction of aspect and modal auxiliaries. *Linguistics and Philosophy*, 32(3):279–315.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

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.

Manfred Krifka, Carlson Gregory N Pelletier, Francis Jeffry, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. 1995. Genericity: An introduction. *The generic book*, pages 1–124.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 Task 5: QA TEMPEVAL - Evaluating Temporal Information Understanding with Question Answering. In *9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 46–54, Denver, Colorado.

Sharid Loaiciga, Thomas Meyer, and Andrei Popescu-Belis. 2014. English-French Verb Phrase Alignment in Europarl. In *Proceedings of LREC 2014*.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Thomas A. Mathew and Graham E. Katz. 2009. Supervised categorization for habitual versus episodic sentences. In *Sixth Midwest Computational Lingustics Colloquium*, Bloomington, Indiana. Indiana University.

Thomas A. Mathew. 2009. Supervised categorization for habitual versus episodic sentences. Master's thesis, Faculty of the Graduate School of Arts and Sciences of Georgetown University.

Paula Menéndez-Benito. 2012. On dispositional sentences. *Genericity*, 43:276.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? annotating the senses of english modal verbs. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pages 24–26.

Eric V Siegel and Kathleen R McKeown. 2000. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):595–628.

Carlota S. Smith. 1997. *The Parameter of Aspect*, volume 43 of *Studies in Linguistics and Philosophy*. Springer.

Ian Soboroff and Donna Harman. 2005. Novelty detection: The trec experience. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 105–112, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 149–156. Association for Computational Linguistics.

Benjamin D Van Durme. 2010. *Extracting implicit knowledge from text*. Ph.D. thesis, University of Rochester.

Zeno Vendler, 1957. *Linguistics in Philosophy*, chapter Verbs and Times, pages 97–121. Cornell University Press, Ithaca, New York.

Jennifer Williams and Graham Katz. 2012. Extracting and modeling durations for habits and events from twitter. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 223–227, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jennifer Williams. 2012. Extracting fine-grained durations for verbs from twitter. In *Proceedings of ACL 2012 Student Research Workshop*, pages 49–54. Association for Computational Linguistics.

Nianwen Xue and Yuchen Zhang. 2014. Buy one get one free: Distant annotation of chinese tense, event type, and modality. *Proceedings of LREC-2014, Reykjavik, Iceland*.

Alessandra Zarcone and Alessandro Lenci. 2008. Computational models for event type classification in context. In *LREC*.