Automatic Identification of Aspectual Classes across Verbal Readings

Ingrid Falk, Fabienne Martin

Automatic Identification of Aspectual Classes across Verbal Readings

Ingrid Falk and Fabienne Martin
University of Stuttgart
firstname.lastname@ling.uni-stuttgart.de

Abstract

The automatic prediction of aspectual classes is very challenging for verbs whose aspectual value varies across readings, which are the rule rather than the exception. This paper sheds a new perspective on this problem by using a machine learning approach and a rich morpho-syntactic and semantic valency lexicon. In contrast to previous work, where the aspectual value of corpus clauses is determined on the basis of features retrieved from the corpus, we use features extracted from the lexicon, and aim to predict the aspectual value of verbal *readings* rather than verbs. Studying the performance of the classifiers on a set of manually annotated verbal readings, we found that our lexicon provided enough information to reliably predict the aspectual value of verbs across their readings. We additionally tested our predictions for unseen predicates through a task based evaluation, by using them in the automatic detection of temporal relation types in TempEval 2007 tasks for French. These experiments also confirmed the reliability of our aspectual predictions, even for unseen verbs.

1 Introduction

It is well known that the aspectual value of a sentence plays an important role in various NLP tasks, like for instance the assessment of event factuality (Saurí and Pustejovsky, 2012), automatic summarisation (Kazantseva and Szpakowicz, 2010), the detection of temporal relations (Costa and Branco, 2012) or machine translation (Meyer et al., 2013). Since, however, the aspectual value of a sentence results from a complex interplay between lexical features of the predicate and its linguistic context, the automatic detection of this aspectual value is quite challenging.

Studies on the computational modelling of aspectual classes emerged about two decades ago with the work of Passonneau (1988) and Klavans and Chodorow (1992), among others. In probably the most extensive study on the field, Siegel and McKeown (2000) extract clauses from a corpus and classify them into states and events, sorting the latter into culminated and non-culminated events in a subsequent step. The classification is based on features inspired by classic Vendlerian aspectual diagnostics, themselves collected from the corpus. Since, however, these features are collected on a type level, this method does not give satisfying results for verbs whose aspectual value varies across readings (henceforth 'aspectually polysemous verbs'), which are far from exceptional (see section 3)[1].

This problem is directly addressed by Zarcone and Lenci (2008). These authors classify corpus clauses into the four Vendlerian aspectual categories (states, activities, accomplishments and achievements), and like Siegel and McKeown, base their classification on (classic aspectual) features collected from the corpus. However, they additionally employ some syntactic properties of the predicate, a move that enables them to better account for the influence of the linguistic context on the aspectual value of the verb across readings.

Friedrich and Palmer (2014), who extend Siegel and McKeown's (2000) model to distributional features, also address the problem of aspectually polysemous verbs, by making use of instance-based syntactic and semantic features, obtained from an automatic syntactic analysis of the clause.

[1] Type-based classification selects a dominant sense for any given verb and then always assigns it for each reading of this verb.

The approach we present here is designed to tackle the issue of aspectual variability and is complementary to the methods just described. As we know from detailed work on verbal syntax and semantics in the tradition of Dowty (1979), Levin (1993), Rappaport and Levin (1998) and subsequent work, many morpho-syntactic and semantic properties of the verb exert a strong influence on its aspectual value in context. As far as we know, no study on the computational modelling of aspectual classes has tried to systematically take advantage of these correlations between lexical properties and lexical aspect. We aim to capitalise on these correlations with the help of a rich French lexical resource, "Les Verbes Français" (Dubois and Dubois-Charlier (1997; François et al. (2007), henceforth LVF). The LVF is a valency lexicon of French verbs providing a detailed morpho-syntactic and semantic description for each reading (use) of a verb.

Differently from previous work, the instances we classify aspectually are verbal readings as delineated in the LVF (rather than corpus phrases). We therefore study lexical aspect on an intermediate level between the coarse-grained type (verb) level and the fine-grained corpus utterance level. Also, while in previous approaches, the features are collected from corpora, those we make use of are retrieved from the lexicon entries. The substantial advantage of this approach, that heavily makes use of the colossal amount of information manually coded in the LVF, is that it enables us to fully investigate the aspectual flexibility of verbs across readings and the factors that determine it.

For our automatic aspectual classification, we firstly extracted verbal readings from the LVF for a set of 167 frequent verbs chosen in such a way that each of the four Vendlerian aspectual classes are roughly equally represented. A semanticist manually annotated each of the corresponding 1199 readings based on a refinement of the classic Vendlerian 4-way aspectual categorisation. This refinement is motivated by recent studies in theoretical linguistics converging in the view that the traditional quadripartite aspectual typology has to be further refined (see (Hay et al., 1999; Piñón, 2006; Mittwoch, 2013) among many others). Such a refinement enables one to better account for the variable degree of aspectual flexibility among predicates, so as to *e.g.* delineate between 'strictly stative' predicates (e.g. *know*), and those stative predicates that also naturally display an activity reading (e.g. *think*). This annotation provides the gold standard for our classification experiments. For each annotated reading, we then collected morpho-syntactic and semantic features from the LVF, chosen for their relevance for the aspectual value of the verb in context. Based on these features, we trained classifiers to automatically predict the aspectual class of the LVF readings.

We assessed the accuracy of our automatic aspectual classification in a task based evaluation as follows. Costa and Branco (2012) showed that (type-based/verb-level) aspectual indicators improve temporal relation classification in TempEval challenges (Verhagen et al., 2007), which emerged in conjunction with TimeML and TimeBanks (Pustejovsky and Mani, 2003). The tasks involved in these challenges require temporal reasoning. Following Branco and Costa's example, we performed TempEval tasks on the French TempEval data, using aspectual indicators derived from the predictions generated by our classifier. This way, we could show that our aspectual classification based on lexical features is reliable.

The paper is structured as follows. Section 2 presents the resource used. Section 3 explains on which criteria verbal readings were manually annotated. Section 4 describes the features collected from the LVF. Section 5 presents the automatic aspectual classification based on these features. Section 6 presents the aspectual indicators derived from the classification. Section 7 describes how our automatic classification was evaluated through TempEval tasks.

2 The Resource – LVF

The LVF, which roughly covers 12 300 verbs (lemmas) for a total of 25 610 readings, is a detailed and extensive lexical resource providing a systematic description of the morpho-syntactic and syntactico-semantic properties of French verbs[2]. The basic lexical units are readings of the verbs, determined by their defining syntactic environment (argument structure, adjuncts) and a semi-formal semantic decomposition (with a finite repertoire of '*opérateurs*'). Once the idiosyn-

[2]The paper version is available online at `http://talep.lif.univ-mrs.fr/FondamenTAL/`. Online access and electronic versions in XML are available at `http://rali.iro.umontreal.ca/rali/?q=fr/lvf`.

crasies are put aside, this decomposition very roughly uses the same inventory of labels and features as in the lexical templates found in e.g. Pinker (1989) or Jackendoff (1983). In Table 1, we give the sample entries for the verb *élargir* 'widen' to illustrate LVF's basic layout.

Syntactic description (Table 1a). Each reading of a verb is coupled with a representation of its syntactic frames. In principle, a verbal reading can be coupled with a transitive frame (labelled 'T'), a reflexively marked frame ('P') and an intransitive frame ('A', 'N') unmarked by the reflexive. The syntactic description additionally specifies some semantic features of the main arguments (e.g. whether the subject and direct object are animate and/or inanimate, whether the indirect object refers to a location, etc). This information is often crucial for the aspectual value of the reading (e.g. a 'human-only' intransitive frame strongly indicates unergativity and henceforth atelicity).

Semantic description (Table 1b). Each entry in the LVF is also characterised by a semi-formal semantic decomposition providing a rough approximation of the meaning of each verbal reading. Each entry is therefore paired with a finite set of primitive semantic features and labels on the basis of which verbal readings are sorted into 14 semantic classes (*eg. psych-verbs*, *verbs of physical state and behaviour*, etc.). The semantic features and labels used in the semantic decomposition provide other cues about the type of verbs (unergative/ unaccusative verbs, manner/ result verbs, etc.) which is instantiated by each reading. For instance, for the reading 01 of *élargir* 'widen' ('*élargir01*' for short) in Table 1b, 'r/d +qt [p]' roughly corresponds to BECOME(more(p)) ('r/d' stands for '(make) become'; '+qt' stands for an increase along a scale). From this, one can safely infer that *élargir01* is a 'degree achievement' verb.

Derivational properties. The LVF also indicates when a verb is formed through a derivational process, and in the positive case, provides information about the category of the verbal root, thus enabling one to identify deadjectival or denominal verbs. Finally, for each entry is specified which suffix is used for the available reading-preserving deverbal nominalisations and adjectives (*-ment, -age, -ion, -eur, -oir, -ure* or zero-derived nominalisations, and *-able, -ant, -é* adjectives).

3 The annotation

We retrieved 1199 entries (verbal readings) for the selected 167 frequent verbs mentioned earlier. On average, each verb has roughly 15 readings, while 50% have more than 13[3]. These readings were manually annotated according to a fine-grained aspectual classification on a 'telicity scale' of eight values.

At the bottom of the scale are readings that are unambiguously ('strictly') stative (i.e. for which any other aspectual value is excluded), rated with 1 (S-STA). For instance, *élargir02* (see Table 1a) is rated with 1, given (a.o.) its incompatibility with the progressive. Those are distinguished from stative verbs that also display a dynamic reading (e.g. *penser* 'think'), rated with 2 (STA-ACT). Readings that are unambiguously dynamic and atelic ('strict activity' readings) are rated with 3 (S-ACT).

At the top are found achievement readings for which any other aspectual value is excluded, rated with 8 (S-ACH). At the middle of the scale are found 'variable telicity' readings, that have no preference for the telic use in a neutral context and are compatible both with *for-* and *in-* adverbials, rated with 4 (ACT-ACC). For instance, *élargir01* is rated with 4, because (a.o.) it is compatible both with *for-* and *in-* adverbials and has no preference for the telic reading in a neutral context. These variable telicity readings are distinguished from 'weak accomplishment' readings, rated with 5 (W-ACC). Out of context, weak accomplishment readings trigger an inference of completion and have a preference for the telic use; however, they are nevertheless acceptable with a *for*-adverbial (on the relevant interpretation of this adverbial). For instance, *remplir01* 'fill' (*Pierre a rempli le seau d'eau* 'Peter filled the bucket with water') is rated with 5, because it by default triggers an inference of completion, but is nevertheless still acceptable with a *for*-adverbial under the 'partitive' reinterpretation of this adverbial. Under this reinterpretation, described e.g. by Smollett (2005) or Champollion (2013), the sentence triggers an inference of non-completion (Bott (2010), see e.g. *Peter filled the bucket with water for 10 minutes*). 'Strong' accomplishment readings — like *remplir09* (*Cette nouvelle a rempli Pierre de*

[3] Interestingly, the average number of 15 readings per verb very closely matches the number of event categories per verb obtained in the experiment reported by Marvel and Koenig (2015), who propose a new method of automatically categorising event descriptions.

id	frame	encoded information
01	T1308	transitive, human subject, inanimate direct object, instrumental adjunct
	P3008	reflexive, inanimate subject, instrumental adjunct
	A30	intransitive with adjunct, inanimate subject
02	N1i	intransitive, animate subject, prep. phrase headed by *de* (*of*)
	A90	intransitive with adjunct, subject human or thing
	T3900	transitive, inanimate subject, object human or thing

(a) Syntactic descriptions

id	example[a]	semantic decomposition	sem. primitive	sem. class
01	On *élargit* une route/ La route *(s')élargit*.	`r/d+qt large`	become	Transformation
02	Cette veste *élargit* Paul aux épaules/ La robe *élargit* la taille.	`d large a.som`	become	Transformation
03	On *élargit* ses connaissances.	`r/d large abs`	become	Transformation
04	On *élargit* le débat à la politique étrangère.	`f.ire abs VRS`	directed move	Enter/Exit

(b) The four readings illustrated by sample sentences and their semantic description

[a]Literal translations – 01: One widens a road/the road is REFL widened/the road widens. 02: This jacket widens Paul 'at the' shoulders/ The dress widens the waist. 03: One widens one's knowledge. 04: One extends the debate to foreign policy.

Table 1: LVF entries for *élargir*

joie 'This news filled Peter with joy') — are incompatible with the partitive reinterpretation of *for*-adverbials.[4] Those are rated with 6 (S-ACC). Finally, accomplishments that share a proper subset of properties with achievements are rated with 7 (ACC-ACH).

The annotator evaluated each entry with a definite or singular indefinite internal argument, in order to abstract away from the role of the determiner in the aspectual value of the VP (see e.g. Verkuyl (1993)).

We also used a coarser grained aspectual scale and group the verbal readings into the following classes: ATElic (rating 1–3), with VARiable telicity (rating 4), and TELic (5 or more). Table 2 gives an overview of the distribution of the aspectual ratings.

The first finding is that verbs display a considerable aspectual variability across readings, which confirms the need to go beyond the type level for the computational modelling of aspectual classes. The aspectual value of 2/3 of the 151 verbs with more than one reading varies with the instantiated reading (on the 8 value scale). With respect to the coarser grained scale, roughly half of the verbs (82, for a total of 793 readings) have readings in more than one of the three overarching aspectual classes.

[4]The *for*-adverbial is nevertheless compatible with *remplir09*, but only under its (non-partitive) 'result state-related interpretation', under which it scopes on the result state, cf. Piñón (1999); see e.g. *This news filled Peter with joy for ten minutes*.

4 The features

The LVF connects each verbal reading with specific morphological, syntactic and semantic features. Among such features, those that influence the lexical aspect of the verb in context are known to be pervasive: Verbs encoding the BECOME operator in their event structure generally have a telic use; intransitive manner verbs are mostly activity verbs (see e.g. Rappaport Hovav and Levin (1998) and subsequent work); ditransitive verbs like *give* are mostly result verbs (see e.g. Pylkkänen (2008)) and thus accomplishments.[5]. We took advantage of many of these features for our classification. Also, some semantic classes give very clear hints to the lexical aspect of its members. For instance, readings instantiating the class of 'enter/exit verbs' are telic, those instantiating the 'transformation' class are never atelic only, etc.[6] We also made use of features conveyed by the semantic decomposition, in particular its main component (BECOME, DO, ITER, STATE, etc.).

We also took advantage of the encoded information on the suffixes used in reading-preserving nominalisations. For instance, readings with an intransitive but no transitive frame can in prin-

[5]Relevant features are sometimes coded in an indirect way. For instance, the difference between verbs like *donner x à y* 'give x to y', that subcategorise the indirect object, and verbs like *dire x à y* 'say x to y', that do not, is retrievable through the difference in the associated syntactic frames.

[6]On this respect, note that the semantic decomposition of *élargir02*, which involves BECOME, shows the limits of the analysis provided by the LVF: Under the 'spatial' use of which *élargir02* is an instance, degree achievements do not describe events in which an individual undergoes change over time (see Deo et al. (2013)).

1	2	3	4	5	6	7	8
S-STA	ACT-STA	S-ACT	ACT-ACC	W-ACC	S-ACC	ACC-ACH	S-ACH
182	67	175	195	172	227	29	152

(a) 8 value scale

1-3	4	5-8
ATE	VAR	TEL
424	195	580

(b) 3 value scale

Table 2: Aspectual distribution of the 1199 manually annotated verbal readings

Features collected from corpus *Example Clause*	**Related features in LVF**
frequency	–
not or never *She can **not** explain why.*	–
temporal adverb *I saw to it **then**.*	durative adverbial in semantic decomposition
implicit or no external argument *He was admitted to the hospital.*	canonical passive, refl. constr. w. instrumental adj.
past/pres participle *... blood pressure going up.*	–
in adverbial *She built it **in an hour**.*	–
4 tense related features	–
manner adverb *She studied **diligently**.*	manner argument or adjunct
evaluative adverb *They performed **horribly**.*	`+ql` in semantic decomposition
for adverbial *I sang **for** ten minutes.*	durative adverbial in semantic decomposition
continuous adverb *She will live **indefinitely**.*	`+re` (iterative operator) in semantic decomposition

Table 3: Siegel and McKeown's (2000) and LVF features.

Features collected from corpus	**Related features in LVF**
temporal adverbs	temporal arg. or adj.
intentional adverbs	–
frequency adverbs	`+qt` in sem. decomp.
iterative adverbs	`+re` in sem. decomp.
tense	–
only subject	A* or N* frame
presence of direct obj	T* frame
presence of indirect obj	N* frame
presence of locative arg	encoded in frame sem class = L
presence of sent. compl.	encoded in frame
canonical passive	T* and A* schema
subj & dobj, number, animacy, definiteness	plural subj or obj human/animal subj or obj thing subj or obj

Table 4: Zarcone and Lenci's (2008) and LVF features.

ciple characterise unaccusative (telic) or unergative (atelic) verbs. But only the latter undergo *-eur* nominalisation, as in English (see Keyser and Roeper (1984)). The availability of the *-eur* nominalisation is therefore a reliable aspectual feature too.

Tables 3 and 4 compare features used in some previous aspectual classifications and their equivalents in the LVF. As one can check, the LVF features cover most of the features used in Siegel and McKeown (2000) and Zarcone and Lenci (2008)[7]. For obvious reasons, features related to grammatical aspect conveyed by tenses are not covered in our valency lexicon. But overall, our set of features roughly corresponds to those used in previous work, for a total of 38 features.

5 Classifying LVF entries

The items we classified are the 1199 readings for the 167 verbs selected. Our classification task consisted in predicting the right (coarse-grained) aspectual class for these readings (ATE, VAR or TEL). In this supervised learning setting, we applied the classifiers shown in Table 5 with the implementation provided by Weka (Hall et al., 2009), mostly with their default settings[8]. We measured the performance of the classifiers by assessing the accuracy in 10-fold cross-validation, and compared it to the accuracy of a baseline classifier which always assigns the majority class (TEL, `rules.ZeroR`). We also performed a linear forward feature selection using the Naïve Bayes algorithm[9]. This way, nine features were selected, coding, among others:

- the presence of a temporal or manner argument/adjunct in the semantic decomposition;
- the main primitive in the semantic decomposition;
- the use of the suffixes *-ment* and *-ure* in the reading-preserving nominalisation;
- the relative polysemy of the lemma (indicated by the number of its readings);
- a subject that must be inanimate;
- the presence of a reflexive reading.

[7]The features used by Friedrich and Palmer (2014) are mainly derived from those of Siegel and McKeown (2000).

[8]For libsvm (the SVM implementation), we used a linear kernel and normalisation. We selected roughly one classifier from each class.

[9]An exhaustive search with the 38 features in this group was computationally too time-consuming.

Algorithm	*complete*	*selected*
`trees.j48`	61.80	63.00
`rules.jrip`	63.89	61.56
`lazy.kstar`	62.89	**67.47**
`functions.libsvm`	62.72	61.13
`bayes.naivebayes`	60.22	65.80
`baseline`	48.37	48.37

Table 5: Classification accuracy for LVF readings, with *complete* feature set and *selected* in feature selection process.

The results in Table 5 show that the features retrieved from the LVF enable one to predict the aspectual class considerably better than the baseline: The accuracy ranges from 12 points to almost 20 above the baseline accuracy of 48.37. The best configuration, achieving an accuracy of 67.48%, is the `lazy.kstar` classifier based on the feature set reduced by feature selection. A comparison with the results reported in previous work is difficult, due to the great discrepancies in the experimental settings (see the introduction). However, our results clearly show that the aspectual class characterising verbal readings can be predicted with a reasonable precision on the basis of lexical-related information only. They once again empirically confirm the well-documented correlations between lexical aspect and the morpho-syntactic/semantic properties of the verb.

6 Aspectual indicators

In this section, we take a more qualitative look at the results obtained in section 5. We assessed the quality of the predictions of our model (henceforth LVF-model) in two ways. Firstly, we derived *aspectual indicators for the type level*, describing the general 'aspectual profile' of a verb across all its readings. These are later used in the task based evaluation described in section 7[10]. Secondly, we looked at the aspectual values assigned to the readings of particular verbs (see *indicators for the verbal readings* below).

Indicators for the type-level. The aspectual indicators for the type-level are computed on the basis of the aspectual values predicted for each reading of the verb. As shown in Table 6, they are designed to reflect how aspectual values vary across the readings of the verb. For example, the indicator 'v' in Table 6a shows whether there is any variation at all, 't' assesses the presence of at least one telic reading, etc. Whereas the indicators in Table 6a provide qualitative cues, those in Table 6b convey quantitative information. The first three give the proportion of readings of a particular aspectual class. The last three are computed from the probability estimates generated by the `libsvm` classifier.

v.	var	> 1 telicity value for same lemma?
m.	maj	Telicity value of majority
t.	tel	Any telic reading?
a.	ate	Any atelic reading?

(a) Nominal and binary aspectual indicators

1.	%tel	Proportion of telic readings
2.	%ate	Proportion of atelic readings
3.	%var	Proportion of flexible readings
4.	probest.max	Max of probability estimates
5.	probest.min	Min of probability estimates
6.	probest.avg	Average of probability estimates

(b) Numeric aspectual indicators.

Table 6: Aspectual indicators

In order to get an idea of the quality of our predictions, we computed from automatic predictions the aspectual indicators for all annotated verbs. We provide some of them in Table 7 for verbs judged aspectually polysemous by the annotator. These 'automatic' aspectual indicators are given in normal font. For the same verbs, we also computed the 'manual' aspectual indicators, i.e. those computed on the basis of the manual annotations (when possible)[11]. These are set in bold face. The verbs in Table 7a are dominantly telic, those in 7b dominantly atelic and those in 7c dominantly variable. As one can check, the dominant aspectual value is correctly assigned in most cases. Also, in most cases, the proportion of uses of the non-preferred readings closely matches the proportion obtained manually. Unsurprisingly, the sample of verbs predicted to be 'mostly telic' are mostly (quasi-)achievement verbs or strong accomplishments describing 'non-gradual' changes (verbs lexicalising changes involving a two-point scale, e.g. *dead* or *not dead* for *kill*, see e.g. Beavers (2008)). Unsurprisingly again, many verbs predicted to be 'mostly variable' are degree achievement verbs. More remarkably, *remplir* 'fill' is

[10] Assigning a value to the type level was necessary to test our predictions on the TempEval corpus, since aligning each utterance of this corpus with a specific LVF-reading is not feasible.

[11] Indicators derived from the probability estimates are not computable from the manual annotations.

rightly predicted to be 'mostly telic', although it is a verb of gradual change. The model therefore preserves here the crucial distinction between degree achievements associated with a close scale like *remplir*, tolerating atelic readings under some uses although they conventionally encode a maximal point (see Kennedy and Levin (2008)), and achievement verbs associated with an open scale like *élargir* 'widen', that also accept both *for-* and *in-* adverbials, but do not show a preference for the telic reading in absence of any adverbial. These observations suggest that even if predictions for some readings are wrong, the aspectual indicators might still rightly capture the general 'aspectual profile' of verbs at the type level.

Indicators for the verbal readings. We also inspected the predicted values for some predicates and compared them to the values assigned manually. For predicates showing a high degree of aspectual variability like *élargir* 'widen' (see Table 7c), the results are very good: *élargir01* ('They are widening the road') is correctly analysed as VAR and *élargir04* ('They are extending the majority') as TEL. Interestingly, *élargir02* ('This jacket widens Pierre's shoulders') is correctly analysed as ATE, despite of the fact that it is wrongly analysed by the LVF as instantiating the class of change of state verbs (see footnote 6). This suggests that the computational model could leverage the information provided by the syntactic frames associated to *élargir02* (see Table 1b) to outweigh the wrongly assigned semantic class and produce the correct aspectual prediction.

lemma	m	t	a	%tel	%ate	%var
casser	**TEL**	**1**	**0**	**95.00**	**0**	**0.05**
'break'	TEL	1	1	95.65	4.35	0
mourir	**TEL**	**1**	**1**	**75.00**	**25.00**	**0**
'die'	TEL	1	1	75.00	25.00	0
remplir	**TEL**	**1**	**1**	**70.00**	**30.00**	**0**
'fill'	TEL	1	1	80.00	20.00	0

(a) Mostly telic

lemma	m	t	a	%tel	%ate	%var
regarder	**ATE**	**0**	**1**	**0**	**91.67**	**8.33**
'look at'	ATE	1	1	16.67	83.33	0
chanter	**ATE**	**0**	**1**	**0**	**66.67**	**33.33**
'sing'	ATE	0	1	0	66.67	33.33
étudier	**ATE**	**1**	**1**	**30.00**	**60.00**	**10.00**
'study'	ATE	1	1	20.00	80.00	0

(b) Mostly atelic

lemma	m	t	a	%tel	%ate	%var
vieillir	**VAR**	**0**	**1**	**0**	**11.11**	**88.89**
'get older'	VAR	0	1	0	22.22	77.78
embellir	**VAR**	**0**	**1**	**0**	**33.33**	**66.67**
'beautify'	VAR	1	0	33.33	0	66.67
élargir	**VAR**	**1**	**1**	**25.00**	**25.00**	**50.00**
'widen'	VAR	1	1	25.00	25.00	50.00

(c) Mostly variable

Table 7: Aspectual indicators computed from predictions and from manual annotations. Indicators in **bold face** are computed based on manual annotations. The names of the indicators refer to the labels used in Table 6.

7 Task based evaluation

Reliable automatic aspectual classifications are expected to enhance existing solutions to temporal relation classification. Thus, if our LVF-model improves such a solution, we can conclude that our learned aspectual values are reliable. We therefore evaluated the predictive power of the LVF-model on unseen verbs through such tasks, following the method proposed in Costa and Branco (2012). While Costa and Branco (2012) collected their aspectual indicators from the web and improved the temporal relation detection in the Portuguese TimeBank (PTiB), we derive ours from the predictions generated using the LVF-model, as described in section 6 and use them in TempEval tasks for the French TimeBank.

The data used in these experiments are the French TempEval data, a corpus for French annotated in ISO-TimeML (FTiB in the following) described in Bittar et al. (2011). This data contains about 15 000 tokens[12] annotated with temporal relations. Of these, roughly 2/3 are marked between 2 event arguments and 1/3 between an event and a temporal expression. The classification tasks we are concerned with deal with the automatic detection of the type of these temporal relations, namely the tasks A, B and C in the TempEval 2007 challenge[13]. Table 8 gives an overview of the data for each of the three classification tasks. We build our experiments on top of a base system addressing these challenges and show that the performance of this base system can be improved using our aspec-

[12] This corresponds to 1/4 of the English TimeBank.

[13] Task A is about temporal relations between an event and a time, task B focuses on relations between events and the document's creation time, and task C is concerned with relations between two events.

	FTiB		PTiB	LVF	
	tlinks	rel. types	tlinks	lemmas(seen)	readings
A	302	10	1659	164(16)	1597
B	264	5	2887	149(14)	1329
C	1172	15	1993	427(40)	3827

Table 8: Event instances for TempEval tasks A, B and C for French and Portuguese (left) and corresponding verbs and readings in LVF (right).

Attribute	A	B	C
event-aspect	×	✓	✓
event-polarity	×	✓	✓
event-pos	×	✓	✓
event-class	✓	✓	✓
event-tense	×	×	×
event-mood*	×	✓	✓
event-vform*	×	✓	✓
order-adjacent	×	N/A	N/A
order-event-first	×	N/A	N/A
order-event-between	✓	N/A	N/A
order-timex-between	×	N/A	N/A
timex-mod	×	×	N/A
timex-type	✓	×	N/A
tlink-relType	class		

Table 9: Features used in the base system for TempEval tasks A, B and C. Features checked (✓) were selected in the feature selection process.

tual indicators.

Like Costa and Branco (2012), we implemented as base system the classifiers proposed for English by Hepple et al. (2007), which only rely on relatively simple annotation attributes. Table 9 lists the features used in the context of our FTiB data, basically the same as in Hepple et al. (2007) and Costa and Branco (2012). As in their work, we also determined the final set of features by performing an exhaustive search on all possible feature combinations for each task, using again the Naïve Bayes algorithm. The features marked '✓' are those finally selected this way. Using this set of features, we trained the same classifiers and under the same conditions described in section 5 on the FTiB data. The accuracy of the resulting models in 10-fold cross-validation on the three TempEval tasks are shown in italics in Table 10.

Following again Costa and Branco (2012), we then enhanced this basic set of features with each of the aspectual indicators computed from the predictions generated by the LVF-model. The aspectual indicators are listed in Table 6; we described their computation in section 6. This way, we obtained 10 enhanced feature sets, one for each aspectual indicator. Using these feature sets and the same classifiers as before, we learned models on the FTiB data and computed their accuracy in 10 fold cross-validation.

The improvements achieved this way are shown in Table 10. Whenever an aspectual indicator improves the results of the base system, we give its accuracy (in bold face) below the accuracy of the base system. The superscripts refer to the lines in Table 6 and show which of the aspectual indicators was used to enhance the base feature set to obtain the reported improved accuracy[14].

The results given in Table 10 show that the accuracy of 8 out of the 15 tested classifiers could be improved by 1-3 points by adding the aspectual indicators. The indicator which produced the most and largest improvements was the average over the probability estimates, suggesting that this value best reflects the dominant aspectual value of the verb. Overall, the improvement obtained through our classification is quantitatively comparable to the enhancement realised by Costa and Branco (2012): Their results show an improvement similar in size to ours for 9 out of the same 15 classifiers. They evaluate on a test set, whereas we compare accuracy in 10-fold cross-validation. This was necessary since the French TimeBank is considerably smaller (roughly 1/4 of Costa and Branco's data set for Portuguese, see PTiB column in Table 8). As mentioned earlier, a qualitative comparison is nevertheless difficult, given the substantial differences between the data and the methodology used here and there.

The results clearly show however that the LVF-model trained on our annotated lexical entries performs well on unseen predicates.

8 Conclusion and future work

This paper focuses on the issue of aspectual variability for the computational modelling of aspectual classes, by using a machine learning approach and a rich morpho-syntactic and semantic valency lexicon. In contrast to previous work, where the aspectual value of corpus clauses is determined at the type (verb) level on the basis of features retrieved from the corpus, we make use of features retrieved from the lexicon in order to predict an aspectual value for each *reading* of a same verb (as they are delineated in this lexicon). We firstly

[14] We only show improvements of at least 1%, and only show the largest gains in performance.

Classifier	A	B	C
trees.j48	*0.71*	*0.81*	*0.40*
	0.73[1]	**0.82**[3]	**0.41**[a]
rules.jrip	*0.71*	*0.83*	*0.36*
	0.73[6]	**0.84**[6]	**0.37**[m]
lazy.kstar	*0.72*	*0.82*	*0.42*
		0.85[5]	
functions.libsvm	*0.74*	*0.83*	*0.40*
		0.85[6]	
bayes.naivebayes	*0.73*	*0.84*	*0.40*
baseline	*0.72*	*0.62*	*0.29*

Table 10: Accuracy of classifiers obtained on FTiB with base and enhanced feature sets. Values for the base classifiers are in italics. In bold face improvements of an enhanced classifier, no values represent no improvement. Superscripts give the aspectual indicator used to enhance the base feature set and obtain the improved result. They refer to rows in Table 6.

studied the performance of the classifier on a set of manually annotated verb readings. Our results experimentally confirm the theoretical assumption that a sufficiently detailed lexicon provides enough information to reliably predict the aspectual value of verbs across their readings. Secondly, we tested the predictions for unseen predicates through a task based evaluation: We used the aspectual values predicted by the LVF-model to improve the detection of temporal relation classes in TempEval 2007 tasks for French. Our predictions resulted in improvements quantitatively similar to those achieved by Costa and Branco (2012) for Portuguese and thus confirm the reliability of our aspectual predictions for unseen verbs.

The investigation reported here can be further pursued in many interesting ways. One possible line of work consists in exploring the aspectual realisation and distribution of the LVF readings in corpus data. This would also provide means to relate our findings for verbal readings to corpus instances.

Our study strongly relies on the LVF lexical database, a very extensive source of morpho-syntactic and semantic information. For other languages, this kind of information, when it is available, is generally not contained in a single lexicon. Therefore, a further interesting research direction would be to evaluate the applicability of our technique to suitable information from distributed resources. On this respect, recent efforts made for linking linguistic and lexical data and making these data accessible and interoperable would certainly be very helpful. For English in particular, available suitable resources are already abundant. One of these is the *Pattern Dictionary of English Verbs*, see (Hanks, 2008). Other interesting data bases are FrameNet (Baker et al., 1998), VerbNet (Levin, 1993; Kipper Schuler, 2006) and PropBank (Palmer et al., 2005), especially since these different resources have been mapped together by (Loper et al., 2007), thus giving access to both the lexical and distributional properties defining each entry.

Increasing the reliability of automatic identification of aspectual classes also represents interesting opportunities for several NLP applications. A finer-grained and more reliable automatic assessment of aspectual classes can among others be quite useful for increasing the accuracy of textual entailment recognition, and, particularly, the sensitivity of systems to event factuality (Saurí and Pustejovsky, 2009). For instance, for telic perfective sentences, while the inference of event completion amounts to an entailment with strong accomplishments and (quasi-)achievements (at least in absence of an adverb signalling incompletion like *partly*), the same inference is to some extent defeasible with weak accomplishments. Integrating finer-grained distinctions among predicates could also enable one to better disambiguate verbal modifiers like durative adverbials. A *for*-adverbial typically signals that the event is incomplete when it modifies a weak accomplishment; e.g., *Peter filled the truck for one hour* suggests that the filling event is not finished, see (Bott, 2010) a.o. However, the same adverbial does not trigger this inference when it applies to a strong accomplishment or a (quasi)-achievement. For instance, *They broke the law for five days* does not suggest that the breaking event is not finished. A system that performs better in the identification of fine grained aspectual classes would therefore evaluate with more precision the probability that the reported event is completed in the actual world.

Acknowledgments

This research was funded by the German Science Foundation, SFB 732 *Incremental specification in context*, Project B5 *Polysemy in a Conceptual System*. For feedback and discussions, we thank Achim Stein and the reviewers of *Sem 2016.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Pro-*

ceedings of the 17th International Conference on Computational Linguistics, volume 1, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

John Beavers. 2008. Scalar complexity and the structure of events. In John Dölling and Tatjana Heyde-Zybatow, editors, *Event Structures in Linguistic Form and Interpretation*. De Gruyter, Berlin.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French TimeBank: An ISO-TimeML annotated reference corpus. pages 130–134. Association for Computational Linguistics.

Oliver Bott. 2010. *The Processing of Events*. John Benjamins, Amsterdam/Philadelphia.

Lucas Champollion. 2013. The scope and processing of *for*-adverbials: A reply to Deo and Piñango. In Todd Snider, editor, *Proceedings of Semantics and Linguistic Theory (SALT) 23*, pages 432–452. CLC publications, Cornell University, Ithaca:NY.

Francisco Costa and António Branco. 2012. Aspectual type and temporal relation classification. pages 266–275. Association for Computational Linguistics.

Ashwini Deo, Itamar Francez, and Andrew Koontz-Garboden. 2013. From change to value difference. In Todd Snider, editor, *Semantics and Linguistic Theory (SALT) 23*, pages 97–115. CLC publications, Cornell University, Ithaca:NY.

David Dowty. 1979. *Word Meaning and Montague Grammar : The semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. D. Reidel Pub. Co., Dordrecht; Boston.

Jean Dubois and Françoise Dubois-Charlier. 1997. *Les Verbes français*. Larousse.

Jacques François, Denis Le Pesant, and Danielle Leeman. 2007. Présentation de la classification des Verbes Français de Jean Dubois et Françoise Dubois-Charlier. *Langue française*, 153(1):3–19.

Annemarie Friedrich and Alexis Palmer. 2014. Automatic prediction of aspectual class of verbs in context. volume 2, pages 517–523. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining Software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Patrick Hanks. 2008. Mapping Meaning onto Use: A Pattern Dictionary of English Verbs. Utah. AACL.

Jennifer Hay, Christopher Kennedy, and Beth Levin. 1999. Scalar structure underlies telicity in 'degree achievements'. In Devon Strolovitch Tanya Matthews, editor, *Semantics and Linguistic Theory (SALT) 9*, pages 127–144.

Mark Hepple, Andrea Setzer, and Robert Gaizauskas. 2007. USFD: Preliminary exploration of features and classifiers for the TempEval-2007 Task. pages 438–441. Association for Computational Linguistics.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press, Cambridge, Mass.

Anna Kazantseva and Stan Szpakowicz. 2010. Summarizing short stories. *Computational Linguistics*, 36(1):71–106.

Christopher Kennedy and Beth Levin. 2008. Measure of change: The adjectival core of verbs of variable telicity. In Louise McNally and Christopher Kennedy, editors, *Adjectives and Adverbs: Syntax, Semantics, and Discourse*, pages 156–182. Oxford University Press, Oxford.

Samuel Keyser and Thomas Roeper. 1984. On the middle and ergative construction in English. *Linguistic Inquiry*, 15:381–416.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Judith L. Klavans and Martin Chodorow. 1992. Degrees of stativity: The lexical representation of verb aspect. In *COLING 1992 Volume 4: The 15th International Conference on Computational Linguistics*.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Edward Loper, Szu ting Yi, and Martha Palmer. 2007. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*.

Aron Marvel and Jean-Pierre Koenig. 2015. Event categorization beyond verb senses. In *Proceedings of the 11th Workshop on Multiword Expressions, NAACL 2015*, pages 569–574.

Thomas Meyer, Cristina Grisot, and Andrei Popescu-Belis. 2013. Detecting narrativity to improve English to French translation of simple past verbs. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 33–42, Sofia, Bulgaria, August. Association for Computational Linguistics.

Anita Mittwoch. 2013. On the criteria for distinguishing accomplishments from activities, and two types of aspectual misfits. In Boban Arsenijevic, Berit Gehrke, and Rafael Marín, editors, *Studies in the Composition and Decomposition of Event Predicates*, pages 27–48. Springer, Berlin.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Rebecca J. Passonneau. 1988. A computational model of the semantics of tense and aspect. *Computational Linguistics*, 14(2):44–60.

Christopher Piñón. 1999. Durative adverbials for result states. In *Proceedings of the 18th West Coast Conference in Formal Linguistics*, pages 420–433. Cascadilla Press, Somerville, MA.

Christopher Piñón. 2006. Weak and strong accomplishments. In Katalin Kiss, editor, *Event Structure and the Left Periphery. Studies on Hungarian*, pages 91–106. Springer, Dordrecht.

Steven Pinker. 1989. *Learnability and Cognition: The Acquisition of Argument Structure*. MIT Press, Cambridge, Mass.

James Pustejovsky and Inderjeet Mani. 2003. Annotation of temporal and event expressions. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Tutorial Abstracts*.

Liina Pylkkänen. 2008. *Introducing Arguments*. MIT Press, Cambridge, MA.

Malka Rappaport Hovav and Beth Levin. 1998. Building verb meanings. In Miriam Butt and Wilhelm Geuder, editors, *The Projection of Arguments: Lexical and Compositional Factors*, pages 97–134. CSLI Publications, Chicago.

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. 43(3):227–268.

Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

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.

Rebecca Smollett. 2005. Quantized direct object don't delimit after all. In Henk Verkuyl, Henriëtte de Swart, and Angeliek van Hout, editors, *Perspectives on Aspect*, pages 41–59. Kluwer Academic Publishers, Amsterdam.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Henk Verkuyl. 1993. *A Theory of Aspectuality : The Interaction between Temporal and Atemporal Structure*. Cambridge University Press, Cambridge; New York.

Alessandra Zarcone and Alessandro Lenci. 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).