Aspect-based Learning of Event Duration Using Parallel Corpora*

Tanja Samardžić and Paola Merlo

Abstract The automatic identification of temporal properties of events is crucial for building semantic representations of texts. In this paper, we present a model which predicts the typical duration (short vs. long) of an event described by a verb, based on structural markings of verb aspect. The information about verb aspect marking is collected from a corpus of Serbian and transferred to English via a parallel corpus. Our study shows that English native speakers' intuitions about possible duration of events described by verbs can be predicted from the patterns of formal aspect marking in the equivalent Serbian verbs.

1 Introduction

Sentences denote events and states that can last from less than a second to an unlimited time span. While people have intuitions on how long an event can last (Pan et al, 2011), it is not clear how this knowledge can be gathered from natural language. In English, like in many other languages, event duration is not a grammatical category like, for example, tense or person. The time span in which an event or a state takes place is understood mostly implicitly.

Tanja Samardzic
CorpusLab, URPP Language and Space, University of Zurich, e-mail: Tanja.Samardzic@uzh.ch

Paola Merlo
Department of Linguistics, University of Geneva, e-mail: Paola.Merlo@unige.ch

* We thank the editors of this volume for giving us the opportunity to contribute to the memory of an esteemed colleague and friend. It is particularly apt we offer this joint contribution, as it was indeed Adam Kilgarriff who suggested to Tanja, a beginning graduate student looking for supervisors interested in verbs, to go to Geneva and work with Paola. This advice has started a fruitful and enduring collaboration.

In natural language processing, establishing event duration is an important or even necessary component of tasks such as question answering and multi-document summarisation. It is crucial for these tasks to establish the temporal order of events (Lapata and Lascarides, 2006, UzZaman et al, 2013). The ordering of events in a discourse depends, among other factors, on the duration of each event separately. For example, the sentences in (1) represent two different orderings. The fact that both events are interpreted as short in (1a) allows a sequential reading of the two events, and an interpretation where the first event is the cause of the second. This reading is not available in (1b), where the duration of the second event is long, spanning over the duration of the first event.

(1)a. John entered the president's office. The president woke up.
b. John entered the president's office. The clock on the wall ticked loudly.
(Dowty, 1986)

Such distinctions are not explicitly encoded in English. Without observable indicators, they need to be inferred from various clues provided in texts (Costa and Branco, 2013). A common approach to identifying the time value of a sentence is to rely on a range of indicators which do not necessarily determine the duration of an event, but can be more or less directly related to it. Such indicators include the morphological form of the words expressing the event – for instance verb tenses in English – and the context of these words, such as time adverbials, syntactically related words, other words found in the context (Pan et al, 2011, Gusev et al, 2011, Williams and Katz, 2012).

Differently from others, our approach to event duration relies on the grammatical category of *verb aspect*. Verb aspect is a general, cross-linguistic property of verbs related to the internal structure of the events they describe. In particular, verb aspect determines whether an event involves a change or not (dynamics), whether it has an end point or not (boundedness) and how long it lasts (duration).

It is a well-known fact that Slavic languages, unlike most of the other European languages, encode verb aspect through lexical derivations. Based on the assumption that aspectual meaning is preserved across languages at the level of lexical items, we collect the information about aspect encoding from a Slavic language (Serbian) and use parallel corpora to transfer it to English.

Automatic identification of aspectual meaning, however, is not a trivial task even in Slavic languages. Although verb aspect is generally marked in these languages through affixation, the marking is highly ambiguous and inconsistent, presenting numerous challenges for generalisations.

To preview, we have developed a quantitative representation of verb aspect, which is based on the distribution of morpho-syntactic realisations of Serbian verbs in a set of parallel English-Serbian verb instances. We show that verb aspect can be automatically acquired from these lexical derivations and transferred across languages using parallel corpora, in our case from Serbian to English. Based on this transferred aspectual information, we predict event duration in English.

The supervised version of our model equals the performance of much more complex models on out-of-domain data and approaches it on in-domain data, thus con-

firming our working hypothesis that verb aspect is a crucial component in determining the perception of event duration.

2 Related Work

Our work draws on two lines of research: automatic identification of event duration and automatic transfer of linguistic information across languages.

Pan et al (2011) learn binary classification of events into short and long using a set of lexical features (the lexical item which expresses the event, its part-of-speech tag and lemma, its WordNet (Fellbaum, 1998) hypernyms) and a set of features extracted from the context of the event (the words surrounding the event, their lemmas and part-of-speech tags, the syntactic object of the event).

Gusev et al (2011) learn event duration using predefined word patterns. Each pattern signals either a long or a short event. One of the patterns used to signal a short event, for example, is Past Tense + *yesterday*. The occurrence data are extracted from the web. Gusev et al (2011) compare the performance of classifiers trained on manual annotation with those trained on the instances where the event duration annotation is replaced by the pattern definitions. They find that a maximum entropy algorithm reaches the best performance of 74.8% accuracy without significant difference between the two settings.

Williams and Katz (2012) explore other word patterns which indicate event duration for classifying events into habitual and episodic. The data are collected from a corpus of Twitter messages and classified using a semi-supervised approach. The study finds that most verbs are used in both senses (habitual and episodic) and proposes a lexicon of mean duration of episodes and habits expressed by a set of verbs. These temporal quantifications, however, are not directly evaluated against human judgments.

Interestingly, feature analyses by both Pan et al (2011) and Gusev et al (2011) indicate that enriching the models with context information brings little or no improvement to the results. Drawing on these findings, we propose a model for learning event durations that employs only word-level features. Our features are based on a fine-grained analysis of temporal properties of events and their grammatical encoding in verb aspect in Serbian.

In applying Serbian verb aspect representation to events expressed in English, we rely on the general method of cross-linguistic transfer of surface morphological information using parallel corpora introduced by Tsang and Stevenson (2001). We follow Tsang et al (2002) in using this methodology to transfer type-level, rather than instance-level properties.

Several cross-linguistic transfer methods have been proposed for various tasks, such as part-of-speech tagging (Snyder et al, 2008), synonym detection (van der Plas and Tiedemann, 2006), syntactic parsing (Kuhn, 2004, Snyder et al, 2009, Zarrieß et al, 2010), the analysis of predicate-argument structure (Fung et al, 2007, Wu and Palmer, 2011), and machine translation (Collins et al, 2005, Cohn and Lap-

ata, 2007). In these approaches, cross-linguistic information is used to disambiguate words and constructions. Our cross-linguistic transfer differs from these approaches in that we use the information from parallel corpora to classify rather than disambiguate. We transfer linguistic properties which apply to types that are more abstract than lexical items and that correspond to verb aspect classes. Cross-linguistic transfer of verb aspect through parallel corpora is explored theoretically by Stambolieva (2011), but the study does not report on automatic data processing.

Cross-linguistic transfer has also been used to develop language resources for new languages by projecting existing annotation via parallel corpora from rich-resource languages to poor-resource languages (Yarowsky et al, 2001, Hwa et al, 2002, Padó, 2007, Burchardt et al, 2009, van der Plas et al, 2011). In our approach, the direction of the transfer is reversed, flowing from a low-resource language to a high-resource language. This is possible because we use naturally occurring morphological forms in Serbian as an annotation of the abstract category of verb aspect not overtly encoded in English.

3 The Theory of Verb Aspect

It has been argued in the linguistic literature that verbs can be divided into a (small) number of *aspectual classes*, depending on a set of properties of the events which they describe.

Theoretical accounts of verb aspect do not agree on what properties of events exactly are encoded by this grammatical category, but one can distinguish two recurring notions: temporal boundedness (whether an event has an end point or not) and duration (how long an event lasts). Verb aspect is defined by a complex interaction between these two properties and a number of other properties describing the dynamics of an event (whether it involves a change, a process, a result state, multiple participants, among others).

The temporal boundaries to which we refer here are those that are implicit to the meaning of the verb. For example, although the state in (2) is temporally bounded to two hours, the boundary is imposed by the time adverbial.

(2)Winston stayed in the shop for two hours.

The meaning of the verb *stay* itself does not imply that there is a start or an end point of the state described by it. This event is therefore temporally unbounded. In contrast to this, the meaning of a verb such as *wake up* in (1a) does imply that there is a point in time where the action described by the verb is completed. Such verbs describe temporally bounded actions. The difference in the existence of an implicit time boundary in the interpretation of the verbs *wake up* and *tick* is precisely what creates the difference in the interpretation of the event ordering in (1a) and (1b).

Computational approaches to verb aspect are mostly concerned with using elements of the context to detect certain aspectual classes (Kozareva and Hovy, 2011,

Siegel and McKeown, 2000). The work of Siegel and McKeown (2000), for example, addresses the aspectual classes proposed by Moens and Steedman (1988), showing, by means of a regression analysis, that the context indicators which distinguish between dynamic and stative events are different from the indicators which distinguish between culminated and non-culminated events.[2] Siegel and McKeown (2000) also show that it is harder to distinguish between culminated and non-culminated events than between static and dynamic events.

The actual aspectual classes discussed both in computational linguistics and linguistic theory vary depending on the approach. Although a taxonomy of four aspectual types — states, activities, achievements, and accomplishments — known as Vendler's classes (Vendler, 1967) has a long tradition in the linguistic literature, there is little agreement on what aspectual classes there are and how to distinguish them (Pustejovsky, 1995, Rothstein, 2008, Ramchand, 2008, Marín and McNally, 2011).

In our study, we do not adopt any predefined classification, but adopt as a working hypothesis the notion that temporal boundedness and duration are related (Dowty, 1986). It is reasonable to expect that short events are temporally bounded, since it is easier to imagine a time boundary in something that lasts a handful of seconds than in something that lasts a hundred years. Long events can be expected to be more rarely temporally bounded. Note that our expectations are probabilistic. We do not exclude the possibility for a short event to be temporally unbounded and for a long event to be bounded. However, we expect short temporally unbounded events to be less likely than short temporally bounded events and long temporally bounded events to be less likely than long temporally unbounded event. We expect these dependencies to be strong enough so that the duration of an event can be predicted from aspectual properties of the verb that expresses it. In some languages, like Serbian, these aspectual properties are overtly marked morphologically.

4 Aspect Encoding in the Morphology of Serbian Verbs

The inventory of Serbian verbs contains different entries for describing temporally unbounded and temporally bounded events. Consider, for example, the sentences in (3-4). The verbs *kuva-o* in (3) and ***pro***-*kuva-o* in (4) constitute a pair of lexical entries in fully complementary distribution: *kuva-ti* (infinitive form of *kuva-o*) is used for temporally unbounded events, and ***pro***-*kuva-ti* is used for temporally bounded events, called *imperfective* and *perfective*, respectively.

The two verbs in (3–4) are morphologically related. The perfective verb is derived from the imperfective by adding the prefix *pro-*. This case represents the simplest relation between an imperfective and a perfective form. In some cases, a perfective form can be derived by adding a suffix, as in (5). If a suffix is attached to the bare form directly, the resulting form expresses a very short or instantaneous

[2] The notion of a culminated event roughly corresponds to the notion of a temporally bounded event discussed in our study.

Table 1 Basic patterns of Serbian verb derivation. Each form in the first column is associated with a set of aspectual properties (indicated as meaning) and with a grammatical class (second column). The third column shows the duration types which can be expected based on the encoded properties. The capital letters in the parentheses indicate potential duration subtypes.

Verb form	Grammatical aspect	Expected duration
<stem> <infl> e.g. *kuva-ti* in (3) Meaning: do <stem>	Imperfective	Long (H)
<pref> <stem> <infl> e.g. *pro-kuva-ti* in (4) Meaning: complete and specify <stem>	Perfective	Short (F)
<stem> <suff> <infl> e.g. *kuv-nu-ti* in (5) Meaning: do <stem> once, instantaneously	Perfective	Short (E)
<pref> <stem> <suff> <infl> e.g. *pro-kuva-va-ti* in (6) Meaning: do <pref> <stem> continuously or repeatedly	Imperfective	Long (G)

event. In some cases, a suffix can be attached to a prefixed form, as shown in (6). By attaching a suffix to a prefixed form (which is perfective), the verb gets a new imperfective interpretation, ambiguous between progressive and iterative.

(3) Vinston je često kuva-o.
Winston AUX often cooked.
Winston often cooked.
(Basic imperfective)

(4) Vinston je pro-kuva-o čašu vode.
Winston AUX coooked glass water.
Winston boiled a glass of water.
(Prefixed perfective)

(5) Vinston je kuv-nu-o malo vode.
Winston AUX cooked little water.
Winston boiled a little bit of water.
(Instantaneous perfective)

(6) Vinston je pro-kuva-va-o čašu vode (kada je čuo glas).
Winston AUX cooked glass water (when AUX heard sound).
Winston was boiling a glass of water when he heard the sound.
(Secondary imperfective)

The examples listed in this section provide a general picture of how boundedness is morphologically marked in Serbian, rather than an exhaustive description.[3] What is important for our study is the fact that boundedness properties are observable in

[3] There are other patterns of lexical expression of boundedness in Serbian. For instance, some verbs are perfective in their basic form; some verbs do not have the basic form (they are found only

the verb forms, although the relationship between the form and the meaning is not simple. Both the affixes and the suffixes which are used to change verb boundedness are not specialised aspect markers. In addition to changing boundedness, they can specify and modify the meaning of verbs. Nevertheless, the morphological expression of boundedness in Serbian is potentially less ambiguous than time adverbials and other elements of the context, hence more helpful in determining event boundedness. Moreover, morphological marking, unlike context, is observable in almost all verb uses.

The described perfective and imperfective derivations can potentially encode numerous boundedness and duration classes. For this affixation system to be useful in encoding event duration, it must be the case that affixes can describe distinct duration classes, that these classes can define a total order of durations and that this classification of duration types is shared by English and Serbian. Table 1 shows the relationship between derivational patterns of verbs and the duration of the events that they describe.

There are potentially two kinds of long events. The secondary imperfectives (example 6 and also the fourth row in Table 1) do not have the same temporal properties as the starting, basic imperfective (example (3) and the first row in Table 1). The secondary imperfective contains the resultative meaning introduced by the prefix (Arsenijević, 2007), while the bare form does not. Resultatives are telic and bounded. As a consequence, the meaning of the secondary imperfective is more specific and more anchored in the context. This distinction might prove relevant for event duration. We can expect prefixed imperfective verbs to describe shorter events than basic imperfective verbs (G $<$ H).

Two kinds of short events can be encoded by the combinations shown in Table 1. The prefixed forms (second row and example 4) are expected to express longer events than the suffixed forms (third row and example 5), since the perfective suffix is specialised for very short or instantaneous events, clearly bounded. The order of the types of short events is thus E $<$ F.

Given that the short events are shorter than the long events, the four event types are organised in a total order of duration E $<$ F $<$ G$<$ H. Note that the complexity of the forms is not proportional to the duration types. The most complex forms express the events which are situated in the middle of the duration scale (F and G), while the simpler forms express events on the two ends of the scale.

Based on the analysis presented in this section, we define two verb aspect features *Pf* (for prefix) and *Sf* (for suffix) that, together with the target event duration, constitute our model for predicting event duration. Based on the boundedness of the event they describe, different combinations of the values of the two features define aspectual classes implicitly. These combinations, in turn, define event duration classes.

Note that our approach to Serbian verb derivations relies only on the observations concerning the form of the verb types. The relation between the observable

with prefixes); some perfective verbs have no imperfective counterparts and vice versa. These are less common cases which are not described here in detail, but which are taken into account in our model.

morphology and the hypothesised verb aspect semantics is tested indirectly by predicting event duration. A semantic analysis could provide additional evidence for the hypothesised aspectual classes. The specificity of the different forms, for instance, can be established directly by comparing the positioning of verb derivations in a semantic vector space (Herbelot and Ganesalingam, 2013). Other potential approaches combining morphological and semantic evidence can potentially lead to a better empirical definition of aspectual classes.

In the following sections, we describe in more detail the model that represents the relations between these two features and how the values of the two features are defined and assigned to English verbs.

5 Learning Event Duration with a Statistical Model

The main goal of our model is to determine if the grammatical notion of temporal boundedness in verb aspect encodes the real-world duration properties of events. The model is a formalisation of our hypothesis that the implicit time boundary in the meaning of a verb and the duration of the event described by it are related, as shown in Table 1.

5.1 The Model

The model takes as input the two-feature representation of verb aspect described in Section 4 above and described in more detail in subsection 6.2 below and generates a probabilistic distribution over event duration classes of verbs in English. It consists of three variables:

D for *Duration* encodes the information about event duration. It can take the values "short" and "long".

Pf for *Prefix* encodes the probability that an English verb is word-aligned to Serbian verbs which have a prefix. This probability can mean two things: a) the event described by the English verb is temporally bounded and specified (the second row in Table 1), or b) the event is unbounded, but specified and potentially short, which is the case when the form contains a suffix too (the fourth row in Table 1).

Sf for *Suffix* encodes the probability that an English verb is word-aligned to Serbian verbs which have a suffix. This probability means two opposite things: a) the event described by the English verb is temporally unbounded, but specified (this is the case when the suffix is added to derive the secondary imperfective, as shown in Example (6) and last row of Table 1), and b) the event is temporally bounded and very short, which is the case when the suffix is added to the bare form directly (see Example 5 and third row of Table 1).

We formalise the described relationships between the variables in the model by means of a Bayesian net, shown in Figure 1. We assume that *Sf* depends both on

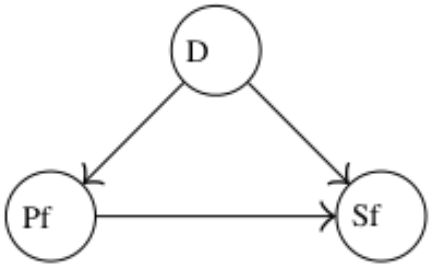

Fig. 1 Bayesian net model for learning event duration

D and *Pf*, which represents the fact that a suffix can be added to a prefixed verb form as a means of deriving secondary imperfectives (potentially expressing long events) or it can be added to a bare form, and it can result in a perfective (expressing a short event). *Pf* depends only on *D*, meaning that a prefix is attached only to the verbs which express events with particular durations (short events). The variable whose values we predict in the experiments is *D*, and the predictors are the other two variables.

The model does not include any lexical information. While English word types are used to collect the aspect features from parallel corpora, as described in Section 6.2, the model itself does not use the information about lexical entries either of English or of Serbian verbs. The input to the model is illustrated with the last three columns in Figure 2.

An important feature of our model is its simplicity. A fine-grained theoretical account of how temporal properties of events are linguistically encoded allows us to make predictions based on only two surface features and to establish a clear relationship between the linguistic theoretical background and our statistical approach. It should be noted here explicitly that this simplicity is voluntary and that it is not simplistic. More complex models have been tested and found not to yield better performance on this task. In particular, a model where the notion of aspect is explicitly represented as a fourth random variable and whose probability was estimated using a hand-annotated Serbian corpus was not found to yield better performance, despite the richer model and the more richly annotated data. The section on results will also compare our method to other more complex methods and show competitive results.

5.2 The Bayesian Net Classifier

We build a classifier which is an implementation of our Bayesian net model described in Section 5.1. Assuming the independence relationships expressed in the Bayesian net (Figure 1), we can decompose the model into smaller factors and cal-

culate its probability as the product of the probabilities of the factors, as shown in (1).

$$P(D,Pf,Sf) = P(D) \cdot P(Pf|D) \cdot P(Sf|Pf,D) \quad (1)$$

The probability of each factor of the product is estimated as the relative frequency of the values of the variables in the training set. The predicted duration value is the one which is most likely, given the values of the two aspect attributes.

$$d^* = argmax_d \;\; P(d|pf,sf) \quad (2)$$

We implement two versions of the classifier: supervised and unsupervised. While the values of all three variables are used to estimate the model parameters in the supervised version, in the unsupervised version, D is treated as an unobserved variable and the parameters are estimated using the expectation-maximisation algorithm.

6 Collecting Verb Aspect from Parallel Corpora

Our model is applied to morphological variables of Serbian to classify event durations in English. This approach requires a word-aligned parallel corpus with the morphological features encoding aspectual information on the source side of the corpus and annotated event duration on the target side of the corpus. Such a corpus, however, does not exist. The examples of annotated event duration are available in English, but we do not have Serbian translations of these sentences. We thus collect the two aspectual features using a different parallel English-Serbian corpus and then combine this information with the English event duration annotation. The resources and methods used to construct our data set are described here.

6.1 Parallel Corpus

The crucial intuition of our method is that aspect information that is implicit in English is more explicitly expressed in Serbian. To transfer verb aspect classification from Serbian to English, we need to know which Serbian verb is the translation of an English verb in a given context. We obtain this information from word alignments in a parallel corpus. In our current study, we use the Serbian translation of the novel "1984" by George Orwell, in the MULTEXT-East project (Krstev et al, 2004, Erjavec, 2010). We obtain word alignments using the system GIZA++ (Och and Ney, 2003), with English as the target language.

We use this corpus for the accuracy of the manual part-of-speech annotation and because the literary text genre is known to be rich in verbs (Biber, 1991). In principle, our methods are applicable to all available parallel corpora. Part-of-speech

tags can be added automatically using automatic tools (Gesmundo and Samardžić, 2012a,b).

In the current study, we transfer the temporal information across languages using parallel corpora, as this is the most straightforward method and parallel resources are available for many languages. If such resources were not available, the transfer could be realised using other methods for establishing word-level translational equivalence. A whole range of such methods have been proposed since the first similarity-based approach of (Rapp, 1995): unsupervised methods based on orthographic, temporal and frequency properties (Schafer and Yarowsky, 2002), or based on decipherement techniques (Ravi and Knight, 2011), or also supervised combinations of multiple monolingual signals (Irvine and Callison-Burch, 2013). Dictionaries obtained from non-parallel corpora do not reach the quality of those obtained from parallel corpora, but they can still be useful in our approach as the transfer is rather resistant to noise.

6.2 Numerical Values of Aspect Attributes

In our cross-linguistic representation, aspectual properties of each English word type are defined by two attributes, *Pf* and *Sf*, whose values vary between 0 and 1. The values are determined based on the observations made in the parallel corpus. They are calculated as the proportion of Serbian verbs containing a particular morphological component out of all Serbian verbs aligned with an English word, as shown in (3).

$$att(e) = \frac{C(algn(e,s), att)}{C(algn(e,s))} \tag{3}$$

where $C(\cdot)$ is a counting function, $att \in \{pf, sf\}$, e is an English word, s is a Serbian verb, $algn(e,s)$ is an aligned English-Serbian pair.

In other words, for each word type on the English side of the corpus, we collect the number of times it is aligned with a prefixed Serbian verb ($C(algn(e,s), pf)$ and we collect the number of times it is aligned with a suffixed Serbian verb ($C(algn(e,s), sf)$.

We also collect counts at the lemma level. For each lemma on the English side of the corpus, we collect the sum of the alignments of all its word types with a perfective Serbian verb. We collect the counts at the two levels because aspectual classes assigned to word types separately are expected to be more precise than assigning the same class to all the forms of one lemma. Summing up the counts for each lemma, on the other hand, is useful as a back-off for classifying word types which are not observed in the parallel corpus.

We do not set any threshold to the number of observations that are included in the measures. We calculate the values for all English types (or lemmas) which are observed at least once in the parallel corpus. To deal with low frequency items and zero counts, we apply add-one smoothing (Russell and Norvig, 2010).

We identify the derivational components of Serbian verbs (prefixes and suffixes) automatically using our own rule-based analyser which implements the linguistic description of lexical derivations, described in Section 4. We do not use the form of Serbian prefixes and suffixes, we only observe whether *any* affix appears in a verb or not.

6.3 Constructing the Data Set

The feature values collected from the parallel corpus are then applied to the word types annotated with event durations in the TimeBank corpus.

Event Durations in TimeBank

The TimeBank corpus (Pustejovsky et al, 2003) is annotated with event durations and other temporal properties of events. The annotation of duration is, in fact, added to a portion of over 2'000 event mentions already annotated in TimeBank, as described by Pan et al (2011). An example of an annotated instance is shown in (7).

(7) There's nothing new on why the plane <EVENT eid="e3" class="OCCURRENCE" **lowerBoundDuration="PT1S" upperBoundDuration="PT10S"**>exploded</EVENT>.

The string that expresses the event (in this case *exploded*) and the event type ("OCCURRENCE") were already identified prior to annotating the duration. In annotating event duration, the annotators were asked to assess the possible time span of an event by determining the span's lower and upper bound (marked with boldface in (7)). Annotators used seven time units (second, minute, hour, day, week, month, year). To obtain a standardised measurement unit, the different units were all converted into seconds in our data. To account for the fact that the perceived difference between time spans decreases as the time span increases, the values in seconds were converted into their natural logarithms. Then the mean value (between the lower and upper bound) was taken as a single duration value. For example, the duration value of the event in (7) is $d = \frac{(ln(1)+ln(10))}{2} = 1.15$. Table 2 illustrates the extracted data set. The first three columns are extracted from the TimeBank.

The values on the logarithmic scale were then used to define a threshold and divide all events into two classes: those that were assigned a value less than 11.4 (which roughly corresponds to a day) were classified as *short events* and the others were classified as *long events*.

After converting these fine-grained annotations into two classes, *short* and *long*, Pan et al (2011) report a proportion of agreement between the annotators of 87.7% on the two classes, which corresponds to a κ-score of 0.75.

Type-based Data Construction

We illustrate the resulting aspectual definitions of English words with a sample of data in Table 2.

		Training/test set		
Type	Lemma	Duration	Pf	Sf
discounts	discount	13.62	NA	NA
talking	talk	13.10	0.9	0.1
boom	boom	19.57	0.5	0.2
estimates	estimate	7.84	0.7	0.3
falling	fall	15.47	0.5	0.1
going	go	17.17	0.4	0.4
carried	carry	8.54	0.5	0.1
fall	fall	4.35	0.6	0.2
falling	fall	15.68	0.5	0.1

Table 2 Example of the data set used in the experiments. Only the last three columns shown here are used for training and testing the model: the *Pf*-value and the *Sf*-value of the word type (collected from Multext-EAST), and the mean duration of the event (annotated in TimeBank). The first two columns illustrate the word type expressing the event and the corresponding word lemma, used to collect the data. They are not used for classification.

There are three main points to note about the collation of the data collected from the parallel corpus and the data in TimeBank. First, our approach is type-based and not token-based.[4] Each instance of a word type in TimeBank is assigned the same pair of values from the parallel corpus (those calculated for that type). It can be seen in Table 2 that the example data set contains two instances of the type *falling* extracted from TimeBank. The two instances are assigned slightly different duration in the manual annotation. In our approach, the two instances are assigned the same pair of *Pf* and *Sf* values (0.5 and 0.1). Second, different word types can be assigned the same pair of values for aspectual features. In this way, lexical items with the same aspectual features are grouped together in aspectual classes. On average, one aspectual class includes two word types. This is illustrated with *falling* and *carried* in our example data set. These two types are assigned the same *Pf* and *Sf* values (0.5 and 0.1). Since the identity of the lexical items expressing the events is not known to the model, these two verb types represent a single class. Third, the value of our attributes can be determined only for those word types which are aligned with Serbian verbs. Another approach is required for the remaining types. We address this issue in Section 7.

[4] In some of the first pieces of work using a similar transfer of morphological information, Tsang and colleagues showed that type-based information transfer is useful for classification (Tsang and Stevenson, 2001, Tsang et al, 2002).

7 Experiments

We present here some experiments that validate the two main aspects of our model. First, we show that, contrary to common belief and practice, events expressed by non-verb word forms have different, and much more easily predictable, duration properties than events expressed by verbs. This justifies our approach to predict event duration based on verb aspect, which applies only to verbs. Second, we also validate the independence assumptions of our model.

We compare our results to other approaches, showing that they are competitive, despite the simplicity of our model. Finally, while we use the TimeBank data set (Pan et al, 2011) for evaluation, we also show that the rather arbitrary setting of the threshold for dividing numerical event durations into short and long in the gold standard can have important consequences for the evaluation of the models. While we adopt the threshold set to one day by Pan et al (2011) to compare to other models, we discuss the effect of setting the threshold with a data-driven method.

7.1 Training and Test Sets

We evaluate the predictions of the model presented in Section 5.1 by training and testing the classifiers described in Section 5.2 on the data set described in Section 6.3.

The TimeBank corpus annotated with event durations consist of 2279 instances. Following Pan et al (2011) and Gusev et al (2011), we create the in-domain portion of the corpus by excluding the Wall Street Journal instances (N=147). We use the train/test split suggested by Gusev et al (2011) for in-domain evaluation. This split is document-based, so that the test set does not contain event mentions from the documents used for training. In addition to this test set, we evaluate our method on the Wall Street Journal portion of the corpus.

As discussed in Section 7.2 below, we divide both the train and the test sets into two portions. We use our model to classify finite and word-aligned instances (671 instance for training, 205 for in-domain testing, 71 for WSJ testing). We classify the remaining instances in the test sets (264 and 76 respectively) by applying the majority class to all of them.

7.2 The Lexical Category of Event Mentions

Events are typically expressed by verbs in a finite form. Predication (establishing a relationship between two entities) is one part of the structure of events which can be expressed by non-finite forms of verbs, by nouns (mostly deverbal), adjectives and other categories too. It is not clear, however, if temporal properties of an event, such as boundedness and duration, can also be expressed by non-finite forms.

Table 3 Percent accuracy of event classification by type of event category, domain type and threshold.

Classifier	Accuracy					
	In-domain test set			WJS test set		
	Finite N=205	Non-finite N=264	Total N=469	Finite N=71	Non-finite N=76	Total N=147
Threshold 7.2						
BNet-S	78.8	81.1	**80.7**	80.3	85.5	83.0
BNet-U	77.6	81.1	79.5	81.7	85.5	**83.7**
Majority class	62.9	81.1	73.1	52.1	85.5	67.3
Threshold 11.4						
BNet-S	73.4	68.9	70.8	76.1	73.7	74.8
BNet-U	65.8	68.9	67.6	73.2	73.7	73.4
Majority class	51.2	68.9	61.2	60.6	73.7	57.1

In the TimeBank annotation, instances of different predicating categories are considered event mentions. The annotation includes nouns (such as *explosion*), adjectives (such as *worth*), and others. A closer look into the manual annotation of event duration in the TimeBank corpus reveals, however, a strong bias towards classifying events as long when event mentions are non-finite. It can be seen in Table 3 how the size of the majority class varies depending on whether event mentions are finite or non-finite. Note that the difference is especially pronounced in the WSJ test set.

This difference has clear practical implications. Always assigning the majority class would thus yield much better accuracy score on non-finite than on finite instances. Based on these observations, we train (and test) our model on the finite event mentions, and assign the majority class to the non-finite ones. The bottom panel of Table 3 shows that our classifier outperforms the baseline in all settings, unsupervised and supervised.

7.3 Validating the Independence Assumptions

Table 4 Percent accuracy of models with different independence assumptions, threshold 11.4.

Classifier	Classification accuracy	
	In-domain set Finite N= 205	WJS set Finite N=71
BNet-S	73.4	76.1
Naive Bayes	72.2	76.0
N-grams	71.7	69.0
Baseline	51.2	60.6

The simplicity of the model raises the question whether the linguistically-justified independence assumptions are needed. The results of the relevant experiment are

shown in Table 4. They demonstrates the appropriateness of the independence assumptions in the model, in comparison to Naive Bayes, a model where the attributes are fully independent, and to a n-gram model where all the attributes are dependent. The n-gram model just counts the n-grams applying back-off smoothing. Both these alternatives yield worse performance than our model.

7.4 Comparison to Other Models

Table 5 Percent accuracy of the unsupervised (U) and supervised (S) version of our classifier (BNet), in comparison with other approaches; threshold 11.4.

	Classification accuracy	
Classifier	In-domain set	WJS set
BNet-U	67.6	73.4
Gus-U1	70.7	73.5
Gus-U2	72.4	**74.8**
BNet-S	70.8	**74.8**
Gus-S	73.0	**74.8**
Pan-S	73.3	73.5

Table 5 shows comparative percent accuracy results of the unsupervised (U) and supervised (S) version of our classifier (BNet), in comparison with two unsupervised approaches proposed by Gusev et al (2011) and supervised approaches by Gusev et al (2011) and by Pan et al (2011), which are described in Section 2. It should be noted that these are very strong competitors, based in one case on very large, Web-scale data collections. Our performance on the in-domain test set approaches the other methods, without quite reaching them. On the out-of-domain WSJ test set, our performance is similar to Gusev et al (2011) in the unsupervised setting. We reach identical score as Gusev et al (2011) in the supervised setting. In this setting, we reach the best performance due to the sharper distinction between finite and non-finite events. Note that all the other approaches also show better results on this set than on the in-domain set.

While we do not outperform but only reach the best existing score in the most challenging setting, these results clearly show the power of strongly linguistically-informed approaches. First, our model is much simpler than the others. It consists of only two word-level features which are theoretically justified. Second, the parameters of the model require very little data to estimate. The model is trained on only a few hundred examples, without need for external resources, such as syntactic parsers or ontologies, which are necessary in other approaches. Third, our approach is more cross-linguistically valid than the others. While the other approaches are language-specific, with features tailored to work for English, applying our approach to another language does not require any adaptation.

One interesting observation is that verb sense does not appear to make a difference. Because of the procedure we use, we learn based on verb aspect types, without any information about specific verbs, but then classify on tokens per verb. Hence, each instance of the same type is assigned the same class. If human judgements varied with the context, the agreement between our classification and human judgments would be low. However, this does not seem to be the case. This is in accordance with the finding of Pan et al (2011) that the best performing setting does not use surrounding context words. It turns out, once again, that "word senses exist only relative to a task" (Kilgarriff, 1997). Identifying deviations from "normal uses" (Hanks, 2013) in the expressions of event duration would require dealing with a broad context.

7.5 Setting the Threshold

It is important to notice that the manual annotation of event durations in the TimeBank corpus did not originally include the distinction between short and long events. The annotators provided their intuitions about the lower and the upper bound of an event duration. These values were subsequently transformed into binary values (see Section 6.3) by manually fixing a threshold. We have followed Pan et al (2011)'s threshold in dividing all the events into short and long, but we show here that a threshold found by a data-driven procedure maximises the results at a different point. This raises the issue whether undue importance in establishing the results is given to what appears to be an arbitrary decision.

To find the optimal threshold, we train the unsupervised version of our classifier (see Section 5.2). The EM algorithm used to assess the values of the parameters converges after 15 iterations assigning the label "long" to 80% of instances and "short" to 20%. On the other hand, Pan et al (2011)'s one day threshold (corresponding to the value 11.4 on the logarithmic scale of mean event durations) gives 53% of long vs. 47% of short events. Clearly, while this threshold has the merit of dividing the data into two comparably-sized subsets, it does not correspond to the notion of short and long modelled by our features, and it might therefore not correspond to a notion of short and long that is linguistically relevant. We compare the agreement between the unsupervised classification and the gold standard considering all possible thresholds. The performance of the model is maximised with the threshold set to 7.2, which allows for comparable sizes of the classes. This makes the threshold of 7.2, roughly corresponding to a duration of half an hour, the best threshold for evaluating our model. In other words, our unsupervised model distinguishes between events which are shorter than half an hour and all other events.

Results with this threshold are shown in the top panel of Table 3. We can observe that setting the gold standard threshold to 7.2 maximises the performance of both the unsupervised and the supervised model.

8 Conclusion

We have shown in this paper how morphological encoding can be used to learn temporal properties of events in a simple, cross-linguistically valid way using light resources. Based on the general idea that semantic categories can be more explicitly encoded in some languages than in others, we have designed an approach that can be used to collect the relevant information in one language and transfer it to another. We have collected the information on verb aspect from Serbian and used it to distinguish between expressions of long and short events in English. We have shown that two word-level features transferred across languages can distinguish between short and long events with the same accuracy as much more complex, context-based monolingual sets of features. Given the variety of aspect encoding in the languages of the world, our approach can be extended to other features from other languages, which might prove useful for making other distinction and performing finer-grained classifications.

References

Arsenijević B (2007) Slavic verb prefixes are resultative. Cahiers Chronos 17:197–213

Biber D (1991) Variation across Speech and Writing. Cambridge University Press, Cambridge

Burchardt A, Erk K, Frank A, Kowalski A, Padó S, Pinkal M (2009) Using FrameNet for the semantic analysis of German: Annotation, representation and automation. In: Boas H (ed) Multilingual FrameNets in computational lexicography, Mouton de Guyter, pp 209–244

Cohn T, Lapata M (2007) Machine translation by triangulation: Making effective use of multi-parallel corpora. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, pp 728–735

Collins M, Koehn P, Kučerová I (2005) Clause restructuring for statistical machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Ann Arbor, pp 531–540

Costa F, Branco A (2013) Temporal relation classification based on temporal reasoning. In: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013), Potsdam, Germany, pp 59–70

Dowty DR (1986) The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics. Linguistics and Philosophy 9:37–61

Erjavec T (2010) MULTEXT-East version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta, pp 2544–2547

Fellbaum C (ed) (1998) WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass.

Fung P, Wu Z, Yang Y, Wu D (2007) Learning bilingual semantic frames: Shallow semantic parsing vs. semantic role projection. In: Eleventh Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007), Skovde, Sweden, pp 75–84

Gesmundo A, Samardžić T (2012a) Lemmatisation as a tagging task. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Jeju Island, Korea, pp 368–372

Gesmundo A, Samardžić T (2012b) Lemmatising Serbian as category tagging with bidirectional sequence classification. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)-2012, European Language Resources Association (ELRA), Istanbul, Turkey, pp 2103–2106

Gusev A, Chambers N, Khaitan P, Khilnani D, Bethard S, Jurafsky D (2011) Using query patterns to learn the durations of events. In: IEEE IWCS-2011, 9th International Conference on Web Service, Institute of Electrical and Electronics Engineers (IEEE), Oxford, UK, pp 145–155

Hanks P (2013) Lexical analysis: norms and exploitations. The MIT Press

Herbelot A, Ganesalingam M (2013) Measuring semantic content in distributional vectors. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Sofia, Bulgaria, pp 440–445

Hwa R, Resnik P, Weinberg A, Kolak O (2002) Evaluation translational correspondance using annotation projection. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, pp 392–399

Irvine A, Callison-Burch C (2013) Supervised bilingual lexicon induction with multiple monolingual signals. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, pp 518–523

Kilgarriff A (1997) "I don't believe in word senses". Computers and the Humanities 31(2):91–113

Kozareva Z, Hovy E (2011) Learning temporal information for states and events. In: Proceedings of the Workshop on Semantic Annotation for Computational Linguistic Resources (ICSC 2011), Stanford

Krstev C, Vitas D, Erjavec T (2004) MULTEXT-East resources for Serbian. In: Proceedings of 8th Informational Society - Language Technologies Conference, IS-LTC, Ljubljana, Slovenia, pp 108–114

Kuhn J (2004) Experiments in parallel-text based grammar induction. In: Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, Barcelona, Spain, pp 470–477

Lapata M, Lascarides A (2006) Learning sentence-internal temporal relations. Journal of Artificial Intelligence Research 27:85–117

Marín R, McNally L (2011) Inchoativity, change of state, and telicity: Evidence from Spanish reflexive psychological verbs. Natural Language and Linguistic Theory 29:467–502

Moens M, Steedman M (1988) Temporal ontology and temporal reference. Computational Linguistics 14(2):15–28

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Computational Linguistics 29(1):19–52

Padó S (2007) Cross-lingual annotation projection models for role-semantic information. PhD thesis, Saarland University

Pan F, Mulkar-Mehta R, Hobbs JR (2011) Annotating and learning event durations in text. Computational Linguistics 37(4):727–753

van der Plas L, Tiedemann J (2006) Finding synonyms using automatic word alignment and measures of distributional similarity. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Association for Computational Linguistics, Sydney, Australia, pp 866–873

van der Plas L, Merlo P, Henderson J (2011) Scaling up automatic cross-lingual semantic role annotation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, pp 299–304

Pustejovsky J (1995) The generative lexicon. MIT Press, Cambridge, MA

Pustejovsky J, Hanks P, Saurí R, See A, Gaizauskas R, Setzer A, Radev DR, Sundheim B, Day D, Ferro L, Lazo M (2003) The TIMEBANK corpus. In: Corpus Linguistics, p 647–656

Ramchand G (2008) Verb Meaning and the Lexicon: A First Phase Syntax. Cambridge Studies in Linguistics, Cambridge University Press, Cambridge

Rapp R (1995) Identifying word translations in non-parallel texts. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Cambridge, Massachusetts, USA, pp 320–322, DOI 10.3115/981658.981709

Ravi S, Knight K (2011) Deciphering foreign language. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, pp 12–21

Rothstein S (2008) Telicity and atomicity. In: Rothstein S (ed) Theoretical and crosslinguistic approaches to the semantics of aspect, John Benjamins, Amsterdam, pp 43–78

Russell SJ, Norvig P (2010) Artificial intelligence : a modern approach. Prentice Hall Pearson, Upper Saddle River, N.J.

Schafer C, Yarowsky D (2002) Inducing translation lexicons via diverse similarity measures and bridge languages. In: Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002), Association for Computational Linguistics, Taipei, Taiwan

Siegel EV, McKeown KR (2000) Learning methods to combine linguistic indicators: improving aspectual classification and revealing linguistic insights. Computational Linguistics 26(4):595–628

Snyder B, Naseem T, Eisenstein J, Barzilay R (2008) Unsupervised multilingual learning for POS tagging. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, pp 1041–1050

Snyder B, Naseem T, Barzilay R (2009) Unsupervised multilingual grammar induction. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, pp 73–81

Stambolieva M (2011) Parallel corpora in aspectual studies of non-aspect languages. In: Proceedings of The Second Workshop on Annotation and Exploitation of Parallel Corpora, Hissar, Bulgaria, pp 39–42

Tsang V, Stevenson S (2001) Automatic verb classification using multilingual resources. In: Proceedings of the Fifth Workshop on Computational Natural Language Learning (CoNLL-2001), Toulouse, France, pp 30–37

Tsang V, Stevenson S, Merlo P (2002) Crosslinguistic transfer in automatic verb classification. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002), Taipei, Taiwan, pp 1023–1029

UzZaman N, Llorens H, Derczynski L, Allen J, Verhagen M, Pustejovsky J (2013) SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, pp 1–9

Vendler Z (1967) Linguistics in Philosophy. Cornell University Press, Ithaca

Williams J, Katz G (2012) Extracting and modeling durations for habits and events from Twitter. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Jeju Island, Korea, pp 223–227

Wu S, Palmer M (2011) Semantic mapping using automatic word alignment and semantic role labeling. In: Proceedings of Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, Association for Computational Linguistics, Portland, Oregon, USA, pp 21–30

Yarowsky D, Ngai G, Wicentowski R (2001) Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proceedings of the First International Conference Human Language Technology, San Diego, CA, pp 161–168

Zarrieß S, Cahill A, Kuhn J, Rohrer C (2010) A cross-lingual induction technique for German adverbial participles. In: Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, Uppsala, Sweden, pp 34–42