Zero-shot Cross-Linguistic Learning of Event Semantics

Malihe Alikhani[1]   **Thomas Kober**[2]   **Bashar Alhafni**[*3]   **Yue Chen**[*4]   **Mert Inan**[*1]
Elizabeth Nielsen[*2]   **Shahab Raji**[*5]   **Mark Steedman**[2]   **Matthew Stone**[5]
[1]University of Pittsburgh   [2]University of Edinburgh   [3]New York University Abu Dhabi
[4]Indiana University   [5]Rutgers University

Abstract

Typologically diverse languages offer systems of lexical and grammatical aspect that allow speakers to focus on facets of event structure in ways that comport with the specific communicative setting and discourse constraints they face. In this paper, we look specifically at captions of images across Arabic, Chinese, Farsi, German, Russian, and Turkish and describe a computational model for predicting lexical aspects. Despite the heterogeneity of these languages, and the salient invocation of distinctive linguistic resources across their caption corpora, speakers of these languages show surprising similarities in the ways they frame image content. We leverage this observation for zero-shot cross-lingual learning and show that lexical aspects can be predicted for a given language despite not having observed any annotated data for this language at all.

1 Introduction

Tense and aspect rank among the most ubiquitous, problematic, and theoretically vexed features of natural language meaning (Hamm and Bott, 2018). Systems of tense and aspect differ considerably—but also often quite subtly—across languages. Figure 1 shows how the corpus manifests differences and similarities across languages that align with their grammatical structures. Tense and aspect have received extensive study across cognitive science; see Hamm and Bott (2018). Nevertheless, from a computational point of view, it has been extremely challenging to gain empirical traction on key questions about them: how can we build models that ground speakers' choices of tense and aspect in real-world information? how can we build models that link speakers' choices of tense and aspect to their communicative goals and the discourse context? how can we build models that recognize

tense and aspect? This is particularly challenging because we might have to work with small annotated datasets. The data scarcity issue renders the need for effective cross-lingual transfer strategies: how can one exploit abundant labeled data from resource-rich languages to make predictions in low resource languages?

In this work, we leverage image descriptions to offer new insights into these questions. For the first time, we present a dataset of image descriptions and Wikipedia sentences annotated with lexical aspects in six languages. We hypothesize that across all of the languages that we study, image descriptions show strong preferences for specific tense, aspect, lexical aspect, and semantic field. We adapt the crowdsourcing methodology used to collect English caption corpora such as MSCOCO and Flickr (Young et al., 2014; Lin et al., 2014) to create comparable corpora of Arabic, Chinese, Farsi, German, Russian, and Turkish image captions. We extend the methodology of Alikhani and Stone (2019) to get a synoptic view of tense, lexical aspect, and grammatical aspect in image descriptions in these diverse languages.

Finally, we study the extent to which verb aspect can be predicted from distributional semantic representations across different languages when the model was never exposed to any data of the target language during training, essentially performing zero-shot cross-lingual transfer. We consider predicting lexical aspect at the phrase level an important prerequisite for modelling fine grained entailment relations, such as inferring consequent states (Moens and Steedman, 1988). For example, this is important for keeping knowledge bases up-to-date by inferring that the consequence of *Microsoft **having acquired** GitHub*, is that now, *Microsoft **owns** GitHub*.

Our results show that the grammatical structure of each language impacts how caption information is presented. Throughout our data, we find, as in

---

* Equal contribution.

Arabic	رجل يمشي بجانب الطريق.
	man walking-PRS-MASC-IPFV-3SG nearby street
	A man is walking nearby the street.

Chinese	雙層 公共 汽車 正 在 公路 上 行駛
	double-decker public bus now IPFV road on drive
	Double-decker public buses are driving on the road.

Farsi	اتوبوس‌های دوطبقه در خیابان حرکت می‌کنند.
	bus-PL double-decker in street move do
	Double-decker buses are moving in the street.

German	Zwei Busse fahren an einer Haltestelle vorbei.
	Two buses drive a bus stop past.
	Two buses drive past a bus stop.

Figure 1: An example image from the MSCOCO dataset with Arabic, Chinese, German and Farsi captions. (ID: 000000568439, photo credit: Stephen Day)

Figure 1, that captions report directly visible events, focusing on what's currently in progress rather than how those events must have begun or will culminate. Yet they do so with different grammatical categories across languages: the progressive aspect of Arabic; the unmarked present of German; or the aspectual marker of the imperfective verbs of Chinese describing an event as in progress.

2 Related Work

Linguists and computational linguists have largely focused on aspectuality as it has been used in unimodal communication. Caselli and Quochi (2007) showed how aspectual information plays a crucial role in computational semantic and discourse analyses. Pustejovsky et al. (2010) described how aspect must be considered for event annotations and Baiamonte et al. (2016) incorporated lexical aspect in the study of the rhetorical structure of text. Kober et al. (2020) presented a supervised model for studying aspectuality in unimodal scenarios only in English. In this work however, we focus on image captions that enable us to better understand how humans describe images. We also explore for the first time the potential of zero-shot models for learning lexical aspect across languages and genre.

The field of automatic image description saw an explosive growth with the release of the Flickr30K and MSCOCO datasets (Vinyals et al., 2015). Fewer works however, have studied how humans produce image descriptions (Bernardi et al., 2016; Li et al., 2019). For example, van Miltenburg et al. (2018a) studied the correlations between eye-gaze patterns and image descriptions in Dutch. Jas and Parikh (2015) investigated the possibility of predict-ing image specificity from eye-tracking data and van Miltenburg et al. (2018b) discussed linguistics differences between written and spoken image descriptions. In this work we continue this effort by offering the first comparative study of verb use in image description corpora that we have put together in six different languages. Alikhani et al. (2020); McCloud (1993); Cohn (2013); Alikhani and Stone (2018); Cumming et al. (2017); Alikhani et al. (2019) proposed that the intended contributions and inferences in multimodal discourse can be characterized as coherence relations. Our analyses and computational experiments explore the extent to which different grammatical-based distinctions correlate with discourse goals and contextual constraints and how these findings generalize across languages.

3 Data Collection and Annotation

Given a set of images, subjects were requested to describe the images using the guideline that was used for collecting data for MSCOCO (Lin et al., 2014). The instructions were translated to six target languages. For the Chinese instructions, we reduced the character limits from 100 to 20 since the average letter per word for English is 4.5. Generally, a concept that can be described in one word in English can also be described in one or two characters in Chinese. The original guideline in English as well as the translations can be found in the attached supplementary material.

We recruited participants through Amazon Mechanical Turk and Upwork.[1] All subjects agreed to a consent form and were compensated at an esti-

[1] https://www.upwork.com/

mated rate of USD 20 an hour. We collected captions for 500 unique images (one caption per image in each of the languages that we study in this paper) that were randomly sampled from MSCOCO for each language. The results of our power analysis suggest that with this sample size, we are able detect effects sizes as small as 0.1650 in different distributions of lexical aspect with a significance level of 95% (Faul et al., 2014).

Annotation Effort. The data is annotated by expert annotators for language specific characteristics of verbs such as tense, grammatical and lexical aspect and the Cohen Kappa inter-rater agreements (Cantor, 1996) are substantial ($\kappa > 0.8$) inter-annotator agreement across the languages.

3.1 Methods

To compare captions and text in a different unimodal genre, we randomly selected 200 sentences across all languages from Wikipedia and annotated their lexical aspect. For Arabic, we used MADAMIRA (Pasha et al., 2014) to analyze the image captions which are written in Modern Standard Arabic. We limited the 200 Chinese Wikipedia sentences to 20 characters in length. The word segmentation and part-of-speech tagging are performed using Jieba Python Chinese word segmentation module (Sun, 2012). Traditional Chinese to Simplified Chinese character set conversion was done using zhconv.[2]

The Farsi image captions and the Wikipedia sentences were automatically parsed using *Hazm* library. For German, we used UDPipe (Straka and Straková, 2017) and we have analysed the Russian morphological patterns by pymorphy2 (Korobov, 2015). For Turkish, the morphological analysis of all the verb phrases in the Wikipedia sentences and the captions are performed using the detailed analysis in (Oflazer et al., 1994). While separating noun phrases from verb phrases, stative noun-verbs of existence ("var" instead of "var olmak") were considered as verbs as well, following the analysis by (Çakmak, 2013).

4 Data Analysis

We performed an analysis of our data to study the following questions: What do image descriptions in Arabic, Chinese, Farsi, German, Russian and Turkish have in common? What are some of the

language-specific properties? What opportunities do these languages provide for describing the content of images? In what follows, we first describe similarities across languages. Next we discuss languages specific properties related to tense and aspect.

In general, captions are less diverse as opposed to Wikipedia verb phrases in terms of their verbs vocabulary across the six languages. Table 1 shows the accumulative percentage of top K verbs for the six languages for Wikipedia and image captions. Wikipedia sentences and captions have different distributions of tense, grammatical aspect and lexical aspect across all languages ($p < 0.01$, $\chi > 12.5$). When it comes to Arabic, atelic verbs dominate the verbs used in Arabic captions. However, the stative verbs dominate the verbs used in Wikipedia sentences.

Moreover, present imperfective verbs make 99% and present perfective verbs make 1% of 85 inflected verbs across all Arabic captions. However, this is drastically different in our baseline. Across 200 full Arabic Wikipedia sentences and out of 180 inflected verbs, present perfective and present imperfective make 49.5% and 2% respectively. Whereas, past perfective and past imperfective make 44.6% and 4% respectively.

This largely agrees with what we analyzed for other languages. In the Chinese data, 56% of Chinese caption verbs are imperfective whereas the majority (70%) of the Chinese Wikipedia descriptions are stative. Chinese Wikipedia sentences also have very few atelic descriptions (1.8%) whereas Chinese captions are populated with atelic descriptions. Chinese does not have tense, but we annotated the sentences both in captions and Wikipedia to learn about the number of sentences that present some kind of cues to refer to an event in the past i.e. adverb. In Wikipedia, 26% of sentences refer to events in past but this number decreases to less than 1% in captions. For Farsi, atelic events make up to 72% of Farsi captions and 17% of Farsi Wikipedia. As in Arabic and Chinese, we observed a major difference in distributions of grammatical aspect and tense in Farsi Wikipedia and Farsi captions. Farsi captions are populated with simple and imperfective present verbs. German captions also follow the general trend with 96% of verbs in caption exhibiting imperfective aspect, in comparison to only 57% in Wikipedia. Atelic verbs dominate the Aktionsart distribution of the captions dataset, making up 55%

	Arabic		Chinese		Farsi		German		Russian		Turkish	
	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.
Top 10	0.262	0.688	0.264	0.367	0.364	0.664	0.394	0.582	0.257	0.654	0.283	0.457
Top 30	0.485	0.937	0.396	0.589	0.466	0.854	0.567	0.804	0.455	0.900	0.524	0.666
Top 100	0.832	–	0.650	0.911	0.545	–	0.911	–	0.802	–	0.728	0.856

Table 1: Captions show a limited distribution of verbs in comparison with Wikipedia. Verb use in Chinese and Turkish captions dataset are more diverse than in Farsi and Arabic caption datasets.

of all verb occurrences, whereas only 16% of verbs are atelic in the Wikipedia sample. The trend is conversed for telic verb occurrences, which make up only 4% in the captions dataset, but 43% in the Wikipedia sample. Interestingly, the proportion of stative verbs is roughly equal in captions and Wikipedia.

The Russian data also hold with these general trends: all captions are imperfective, whereas only 50% of Wikipedia sentences are. This distribution is even more extreme in Russian than in other languages partially because of a unique property of the Russian aspectual and tense system: only verbs that refer to past or future events in Russian can be perfective. In the captions, 99% of verbs refer to present events and therefore are required to be imperfective. This also is borne out the telicity of Russian captions: 49% of captions are atelic, 30% are stative, and only 22% are telic. By contrast, only 21% of Wikipedia data is atelic, while 26% is stative, and 53% is telic. As discussed in Section 4.1 below, this reflects a correlation between perfectivity and telicity in Russian.

Telicity of the Turkish data follows a similar distribution to the other languages, with a key difference in the statistics of stative verbs. Both Wikipedia sentences and captions have higher count of stative verbs compared to other languages. 56% of Wikipedia verbs and 63% of caption verbs are stative in Turkish. This is caused by the inherent copula usage and preference of stative and timeless tenses such as the "geniş zaman". Atelic verb percentage in captions (30.4%) is considerably smaller to that of stative verbs (63.8%). There is a drastic difference between the number of telic verbs with a 32.4% in Wikipedia phrases compared to 5.8% in captions.

4.1 Language-Specific Observations

Arabic. Arabic has a rich morphological system (Habash, 2010). Moreover, verbs in Arabic have three grammatical aspects: perfective, imperfective, and imperative. The perfective aspect indicates that

actions described are completed as opposed to the imperfective aspect which does not specify any such information. Whereas the imperative aspect is the command form of the verb.

Similar to German and Russian, non-past imperfective verbs were dominant across the captions in Arabic as opposed to Chinese, Farsi, and Turkish. Furthermore and as shown in Table 2, 72.2% of Arabic captions were atelic, and this is the highest atelic percentage for captions across all languages. Whereas, 8.9% of the Arabic Wikipedia sentences were atelic, which constitutes the lowest atelic percentage for Wikipedia sentences across all other languages. This highlights an interesting evidence of the morphological richness in Arabic and how verbs can inflect for mood and aspect.

Chinese. Chinese is an equipollent-framed language (E-framed language), due to its prominent feature – serial verb construction (Slobin, 2004). For example, 走进 (walk into) and 走出 (walk out of) are treated as two different verbs. This phenomenon greatly enlarged the vocabulary of Chinese verbs perceived by POS taggers and parsers. We believe this is an important reason why Chinese verbs look so diverse and the distribution among atelic, telic and stative looks rather imbalanced. Having the base verb character and adding on aspectual particles changes the telicity. Given the nature of Wikipedia text, it is observed that in table 2 only 1.8% are atelic and more than 69.8% are stative, while in image captions more than 56% are atelic.

Since Chinese does not have the grammatical category of tense, the concept denoted by tense in other languages is indicated by content words like adverbs of time or it is simply implied by context. For example, the verb for "do" is 做 (zuo), which is used to describe all past, present, and future events. Since the verb remains the same, temporal reference is instead indicated by the time expressions (Lin, 2006), for example:

(1) 昨天 我做了 批萨。

215

	Arabic		Chinese		Farsi		German		Russian		Turkish	
	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.
Atelic	0.089	0.722	0.018	0.561	0.171	0.719	0.162	0.550	0.213	0.488	0.114	0.304
Telic	0.371	0.010	0.285	0.063	0.470	0.042	0.431	0.038	0.530	0.218	0.324	0.058
Stative	0.540	0.268	0.698	0.377	0.357	0.237	0.407	0.412	0.257	0.299	0.560	0.638

Table 2: Captions include more atelic descriptions in comparison with Wikipedia across languages.

Yesterday I do PFV pizza.
Yesterday I made pizza.

Farsi. In the Farsi caption dataset four verbs make up to around 50% of the verbs: *to be* (بودن), *to play* (بازی کردن) , *to sit* (نشستن), and *to look* (نگاه کردن)

Table 1 shows difference in verbs distributions across languages. The data regarding the distribution of caption verbs in English are reported by (Alikhani and Stone, 2019). Chinese captions are much more diverse and the difference is statistically significant ($p < 0.05$, $\chi = 14.4$).

Farsi verbs are either simple or compound. Any lexical unit which contains only a verbal root is a simple verb (e.g. verbal root: رفتن 'to go'). The lexical unit which contain either a prefix plus a verbal root, or a nominal plus either a regular verbal root or an auxiliary verb are compound verbs. Related to this is the phenomenon of incorporation, defined by (Spencer, 1991) as the situation in which "a word forms a kind of compound with its direct object, or adverbial modifiers while retaining its original syntactic function."

59.3% of Farsi Caption verbs are compound and 88.2% of the compound verbs are constructed with کردن (to do) and شدن (to be). Wikipedia on the other hand includes only 12.1% compound verbs. Majidi (2011) conjectured that کردن (to do) and شدن (to be) are used when the speaker wants to highlight the meaning of the noun even more in comparison with cases where nouns are accompanied with گرفتن (to take) or داشتن (to have). For example, نگاه کردن (literally *Do a look*) is the fourth most frequent verb in captions.

However, the majority (97%) of the compound verbs in captions are constructed with nouns. Megerdoomian (2002) hypothesized that the aspectual properties depend on the interaction between the non-verbal and the light verb and that the choice of light verb affects argument structure. For instance, to form the transitive version of an intransitive predicate, Farsi speakers replace the light verb by its causative form. **All of the intransitive**

compound verbs in our corpus are atelic.

German. German speakers predominantly used the present simple — rather than the present progressive — to describe atelic activities, where we found that only ≈7% of atelic captions have been described in the present progressive. For example, sentences (1)-(2) below show two captions where the ongoing activity is described in the present simple in German, however in English, the present progressive would be used. In English, the use of the present simple has a strong futurate reading, which is substantially weaker in German. Thus we attribute the frequent use of the present simple in German to it being less aspectually ambiguous.

(1) Zwei Männer **spielen** Wii im Wohnzimmer.

   *Two men **are playing** on a Wii in the living room.*

(2) Ein Mann und eine Frau **fahren** Ski.

   *A man and a woman **are skiing**.*

We furthermore found that German speakers have frequently omitted the verb altogether if an imaged depicted some form of still life. These sentences exhibit stative lexical aspect, and typically, verbs such as "stand", "lie" or a form of "to be" would have been the correct verb as sentences (3)-(4) below demonstrate, where we have added a plausible verb in square brackets.

(3) Ein Zug **[steht]** neben einer Ladeplattform.

   *A train **[is standing]** next to a loading bay.*

(4) Eine Pepperoni Pizza **[liegt]** in einer Pfanne neben einem Bier.

   *A pepperoni pizza **[is lying]** in a pan next to a beer.*

Russian. A distinction between imperfective and perfective aspect must be marked on all Russian verbs. This contrasts with languages (e.g., Spanish) where aspect is only marked explicitly in a subset of the verbal system, such as within the past tense.

Aspect marking in Russian is often done by means of affixation: a default-imperfective stem becomes perfective with the addition of a prefix (e.g. *pisat'* > *napisat'* 'to write' (Laleko, 2008)). Perfective aspect expresses a view of an event "in its entirety" (Comrie, 1976), including its end point, meaning that perfectivity and telicity are highly correlated. For example, the use of the perfective *napisat'* 'to write' implies the completion of a finite amount of writing, whether or not the speaker chooses to include an explicit direct object indicating what is being written. There is disagreement in the literature on whether all perfective verbs in Russian are telic or if the perfectivity is merely correlated with telicity (Guéron, 2008; Filip, 2004). However, the fact that all verbs must be explicitly marked as either perfective or imperfective, combined with the fact that telicity is at least positively correlated with perfectivity, may lead to more verbs in the Russian being labelled as telic. In fact, we do find that when compared with languages such as English, where verbs may remain under-specified for aspect and therefore for telicity, the Russian captions contain significantly more telic verbs.

Turkish. Lexical aspects of verbs in Turkish captions differ from other languages in terms of choice of the sentence structure and the diversity of Turkish tenses, with the presence of copula. These intricacies are analyzed using the work of (Aksan and Aksan, 2006) and (Aksan, 2003). It can be observed that Turkish-speakers tend to choose a specific sentence structure while describing pictures.

Captions are populated with noun phrases consisting of a verbal adjective, a subject and an implicit noun-verb ("var"). The most important aspect about determining lexical aspect in Turkish is the plethora of tenses. A considerably different tense is the "geniş zaman", which translates to "broad time/tense". Its use broadens the time aspect in a verb to an extent that the verb exists in a time-less space. Even though it is generally compared with the present simple tense in English, "geniş zaman" telicity greatly depends on the context and the preceding tense in the agglutinative verb structure. Wikipedia sentences contain 13.3% "geniş zaman" verbs while caption verbs do not have any of that formation. This is due to the difference of giving a description or a definition.

Turkish definitions are timeless and use "geniş zaman" more frequently, while descriptions, like

in the captions, use other tenses. It can be presumed that all "geniş zaman" verbs are atelic; however, this does not necessarily hold true in captions where a limited number of telic cases exist, which increases the importance of a differentiation between atelic and telic tenses in Turkish. Another distinction that is visible between the Turkish image captions and Wikipedia sentences is the progressive aspect. 59.7% of caption verbs are progressive while only 0.9% of Wikipedia verbs are progressive. This aspect is used extensively in captions due to its close relation with any action verb that is being done.

5 Computational Experiments

In this section we leverage our multilingual annotated dataset and investigate to what extent aspect can be detected with computational methods. More specifically, the primary research question we address in this section is an empirical investigation whether distributional semantic models capture enough information about the latent semantics of aspect to be detected across languages.

Our use of distributional semantic representations is furthermore motivated by the fact that they are readily available in numerous languages, and that they, contrary to manually constructed lexicons such as VerbNet (Schuler and Palmer, 2005) or LCS (Dorr and Olsen, 1997), scale well with growing amounts of data and across different languages. Furthermore, there is a growing body of evidence that models based on the distributional hypothesis capture some facets of aspect (Kober et al., 2020; Metheniti et al., 2022), despite the fact that aspect is represented in a very diverse manner across languages.

5.1 Aspectual Classification

We treat the prediction of verb aspect as a supervised classification task and experiment with pre-trained fastText (Grave et al., 2018) embeddings[3], multilingual BERT (Devlin et al., 2019), and ELMo (Peters et al., 2018; Che et al., 2018)[4] as input, and the aspectual classes *state, telic, atelic* as targets. For fastText we average the word embeddings to create a single vector representation, for

---

[3]https://fasttext.cc/docs/en/pretrained-vectors.html

[4]While the BERT model is truly multilingual, we use a single monolingual ELMo model for our experiments from https://github.com/HIT-SCIR/ELMoForManyLangs.

	Aspect	Arabic		Chinese		Farsi		German		Russian		Turkish	
		Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki
fastText	Atelic	**0.95**	-	**0.97**	-	**0.95**	-	**0.90**	-	**0.96**	-	0.51	-
	Telic	-	0.48	-	0.00	-	0.74	-	**0.89**	-	0.83	-	0.62
	State	0.84	0.66	0.00	**0.89**	0.83	0.59	**0.88**	**0.88**	**0.94**	0.27	0.83	0.80
mBERT	Atelic	0.50	-	0.80	-	0.73	-	0.72	-	0.78	-	**0.96**	-
	Telic	-	0.64	-	**0.92**	-	0.75	-	0.84	-	**0.83**	-	**0.79**
	State	0.88	**0.79**	0.91	0.47	**0.93**	0.57	0.82	0.82	0.88	**0.44**	0.91	**0.89**
ELMo	Atelic	0.65	-	0.76	-	0.77	-	0.78	-	0.90	-	**0.97**	-
	Telic	-	**0.66**	-	0.87	-	**0.79**	-	0.76	-	0.83	-	0.74
	State	**0.89**	0.78	0.88	0.22	**0.93**	**0.67**	0.85	0.75	**0.94**	0.20	**0.93**	0.86

Table 3: Mono-lingual F1-scores per label across all languages with using fastText embeddings (top), multilingual BERT embeddings (middle) and ELMo embeddings (bottom).
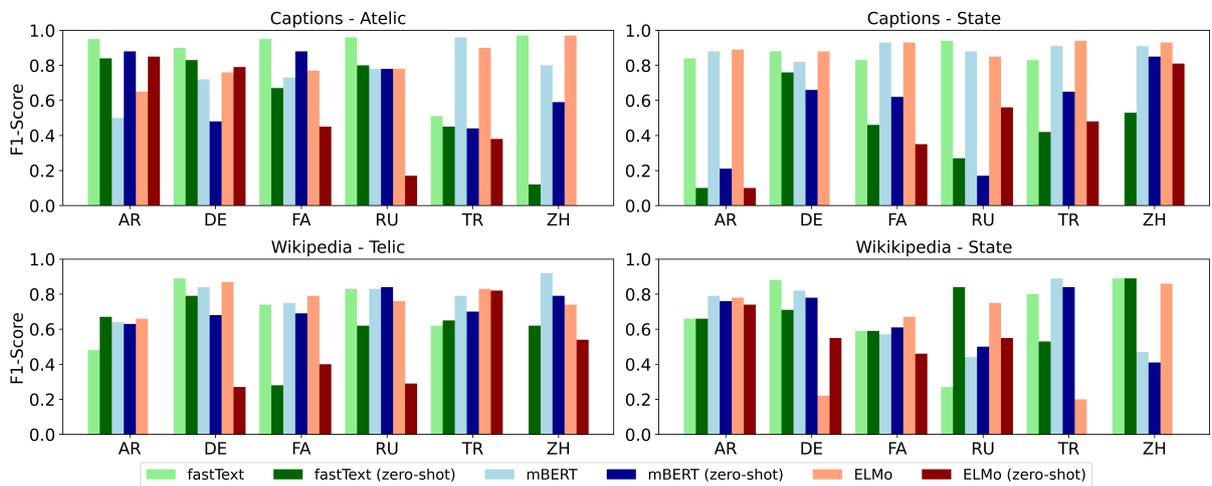

Figure 2: Performance comparison between zero-shot cross-lingual (darker shades) learning and a mono-lingual (lighter shades) setup. Remarkably, even without any target language data, our simple zero-shot setup is competitive with using mono-lingual data and even surpasses it in some cases.

multilingual BERT we use its `[CLS]` token, and for ELMo the pooled representation of the encoded utterance for classification. We use the Logistic Regression classifier from scikit-learn (Pedregosa et al., 2011) with default hyperparameter settings.

Our choice of models is motivated by: a) assessing performance with a word-level model (fast-Text), b) estimating the performance difference when large pre-trained models (ELMo & mBERT) are applied, and c) observing the difference between a single multilingual model (mBERT) and monolingual models for the different languages (fastText & ELMo).

Mono-lingual. For the mono-lingual experiments, we evaluate our method on the annotated captions and Wikipedia sentences, however we decided to drop all *telic* instances from the captions data, and all *atelic* instances the Wikipedia sentences, as they occur very infrequently in either respective corpus.[5] We are focused on establishing whether aspect can be predicted from embeddings across languages *in principle* and wanted to avoid obfuscating the problem of predicting aspect with the problem of class imbalance.

The aim of our first experiment is to establish that aspect can be classified for our set of languages with distributional representations in a supervised setting as has been shown on English data (Kober et al., 2020). Figure 3 shows the difference in Accuracy of our models in comparison to a majority class baseline. As the figure shows, the distributional models are able to outperform the majority

---

[5]This reduced the classification problem to a 2-class problem, *stative* vs. *non-stative*.
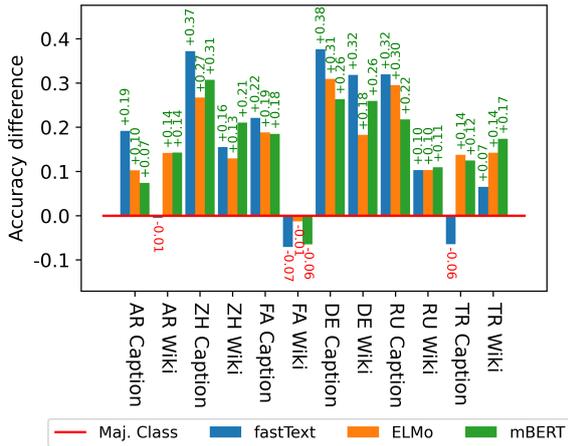
218

Figure 3: Accuracy comparison of a majority class baseline to fastText, multilingual BERT and ELMo models across all our target languages and domains.

class baseline by substantial margins across the board with the exception of our Farsi Wikipedia dataset where we underperform the baseline by a small margin.

Next, we aim to establish baseline scores for the distributional models on our dataset. We perform stratified 10-fold cross-validation on our annotated datasets and report a micro-averaged F1-Score on the basis of accumulating the number of true positives, false positive, and false negatives across all cross-validation runs (Forman and Scholz, 2010).

Table 3 shows that except for Chinese, our simple method of predicting aspect from averaged fastText embeddings works astonishingly well across languages, achieving F1-scores in the mid-80s to mid-90s for many languages. Multilingual BERT and ELMo perform similarly across languages with notable problems for distinguishing between states and *telic* events in Russian and Chinese.

Overall, all models perform approximately in the same ballpark, specifically, there is no dramatic loss in performance when using a single multilingual model in comparison to monolingual models. Conversely, an LSTM-based model and the even simpler bag-of-words based model work remarkably well given the latent nature of aspect. Distributional representations appear to capture enough information for making fine-grained semantic distinctions — an important result for further work on multilingual semantic inference around consequence and causation (Mirza and Tonelli, 2014; Kober et al., 2019; Guillou et al., 2020).

Zero-Shot Cross-lingual. For the zero-shot cross-lingual experiment we use the aligned fast-

Text embeddings and the same mBERT and ELMo models as in the mono-lingual experiments.[6] We perform a zero-shot learning on the basis of a leave-one-language-out evaluation. This means that we train our Logistic Regression classifier on the data of five languages and evaluate performance on the sixth one. The models were never exposed to *any* data of the target language during training, thereby performing zero-shot cross-lingual transfer. This assesses how much information can be leveraged cross-lingually, which has potential further applications for transfer learning and data augmentation.

As for the mono-lingual experiments we drop the *telic* class from the captions data, and the *atelic* class from the Wikipedia data. Figure 2 compares mono-lingual with zero-shot cross-lingual performance, showing that our simple setup yields remarkably strong results, that in some cases even outperform the mono-lingual setup. Our results indicate that a considerable amount of aspectual information can be transferred and induced cross-lingually, providing a very promising avenue for future work.[7] In order to estimate the importance of the contribution of each language in the zero-shot setting we conduct a Shapely-flavoured (Shapley, 1953) analysis. Shapely values are a method for quantifying the contribution to model performance of any given feature in a dataset (Molnar, 2022).

As Shapely values operate on the *feature* space, rather than the *instance* space, we interpret the presence of training data for a particular languages as a binary indicator feature. This means that any languages can be "active" during model training, or not. This way, we can observe the performance of a model with and without any given language in the training data, and estimate that language's impact on model performance. The process to estimate the Shapely-flavoured impact value for a given language is perhaps best explained by an example: supposing our target language — for which we want to predict aspect — is Arabic, and we want to quantify the contribution of German language training data in our model, we start by training a model on Farsi data and compare our model's predictive performance to a model trained on Farsi *and* German data. Next, we train our model on Farsi and Russian data, and compare its performance to a model trained on Farsi, Russian *and*

---

[6]https://fasttext.cc/docs/en/aligned-vectors.html

[7]A multilingual companion table to Table 3 is presented in Table 5 in Appendix 7.
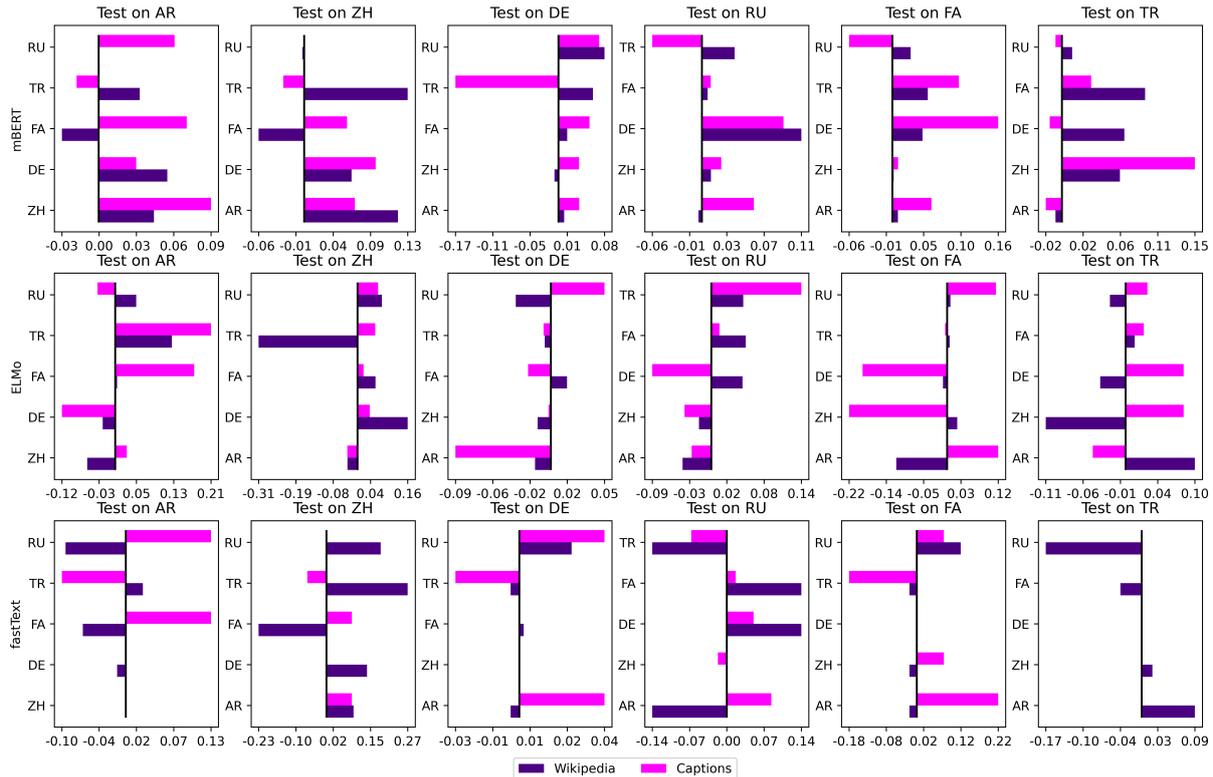
Figure 4: Shapely-flavoured analysis of the impact of each language's presence in the training data on predicting aspect in a target language in a zero-shot cross-lingual setting.

German data, and so on for all combinations of training data. Lastly, we average the differences of all these comparisons to obtain a value that represents the impact of German data on predicting aspect for Arabic. We perform this method for all model, language and domain combinations, with the resulting Figure 4 summarising all Shapely-flavoured impact values for all languages. The figure shows the positive and negative impact of each language — for the captions dataset in magenta and the Wikipedia dataset in indigo — for measuring accuracy. Generally, the impact of each language on model performance is primarily governed by the *kind of model*, rather than the language(s) used for training. While this may seem somewhat dissatisfying at first, we believe that understanding model behaviour is paramount for transfer learning with cross-lingual data with the goal of leveraging e.g. the explicit aspectual markers in the Slavic languages to learn models for languages such as English where aspect is more opaque, as a very fruitful avenue for future research.

6 Conclusion

By analyzing verb usage in image–caption corpora in Arabic, Chinese, Farsi, German, Russian and Turkish we find that people describe visible eventualities as continuing and indefinite in temporal extent. We show that distributional semantic can reliably predict aspectual classes across languages, and achieves remarkable performance even in zero-shot cross-lingual experiments.

Our study has also revealed that these qualitative properties and grammatical differences reflect the discourse constraints in play when subjects write captions for images and that these findings are generalizable across languages. We have leveraged this observation for our computational work where we show that aspect can be predicted with distributional representations in a mono-lingual setup. We have furthermore provided first evidence that aspect can be predicted in a zero-shot cross-lingual manner where a model has not been exposed to any training data in the target language at all.

Acknowledgement

References

Mustafa Aksan and Yeşim Aksan. 2006. Denominal verbs and their aspectual properties. *Dil Dergisi*, pages 7 – 27.

Yeşim Aksan. 2003. Türkçe'de durum değişikliği eylemlerinin kılınış özellikleri. *DİLBİLİM ARAŞTIRMALARI DERGİSİ*.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. Cite: A corpus of image-text discourse relations. *arXiv preprint arXiv:1904.06286*.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Malihe Alikhani and Matthew Stone. 2018. Exploring coherence in visual explanations. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 272–277. IEEE.

Malihe Alikhani and Matthew Stone. 2019. "caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67.

Daniela Baiamonte, Tommaso Caselli, and Irina Prodanof. 2016. Annotating content zones in news articles. *CLiC it*, page 40.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Alan B Cantor. 1996. Sample-size calculations for cohen's kappa. *Psychological methods*, 1(2):150.

Tommaso Caselli and Valeria Quochi. 2007. Inferring the semantics of temporal prepositions in italian. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 38–44. Association for Computational Linguistics.

Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium. Association for Computational Linguistics.

Neil Cohn. 2013. Visual narrative structure. *Cognitive science*, 37(3):413–452.

Bernard Comrie. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, volume 2. Cambridge university press.

Samuel Cumming, Gabriel Greenberg, and Rory Kelly. 2017. Conventions of viewpoint coherence in film. *Philosophers' Imprint*, 17(1):1–29.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving verbal and compositonal lexical aspect for nlp applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 151–158, Madrid, Spain. Association for Computational Linguistics.

F Faul, E Erdfelder, AG Lang, and A Buchner. 2014. G* power: statistical power analyses for windows and mac.

Hana Filip. 2004. The telicity parameter revisited. In *Semantics and Linguistic Theory*, volume 14, pages 92–109.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Jacqueline Guéron. 2008. On the difference between telicity and perfectivity. *Lingua*, 118(11):1816–1840.

Liane Guillou, Sander Bijl de Vroe, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2020. Incorporating temporal information in entailment graph mining. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 60–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Nizar Y. Habash. 2010. Introduction to arabic natural language processing. *Synthesis Lectures on Human Language Technologies*, 3(1):1–187.

Friedrich Hamm and Oliver Bott. 2018. Tense and Aspect. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2018 edition. Metaphysics Research Lab, Stanford University.

Mainak Jas and Devi Parikh. 2015. Image specificity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2727–2736.

Thomas Kober, Malihe Alikhani, Matthew Stone, and Mark Steedman. 2020. Aspectuality across genre: A distributional semantics approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4546–4562, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 103–119, Gothenburg, Sweden. Association for Computational Linguistics.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Oksana Laleko. 2008. Compositional telicity and heritage russian aspect. In *Proceedings of the Thirty-Eighth Western Conference on Linguistics (WECOL)*, volume 19, pages 150–160.

Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. 2019. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312.

Jo-Wang Lin. 2006. Time in a language without tense: The case of chinese. *Journal of Semantics*, 23(1):1–53.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Maryam Majidi. 2011. Lexical aspect in farsi. In *Proceedings of the journal of Persian Literature*, pages 145–158.

Scott McCloud. 1993. *Understanding comics: The invisible art*. William Morrow.

Karine Megerdoomian. 2002. Aspect in complex predicates. In *Talk presented at the Workshop on Complex Predicates, Particles and Subevents, Konstanz*.

Eleni Metheniti, Tim Van De Cruys, and Nabil Hathout. 2022. About time: Do transformers learn temporal verbal aspect? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–101, Dublin, Ireland. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

Kemal Oflazer, Elvan Göçmen, Elvan Gocmen, and Cem Bozsahin. 1994. An outline of turkish morphology.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1094–1101, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

Karin Kipper Schuler and Martha S. Palmer. 2005. *Verbnet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, USA. AAI3179808.

L. S. Shapley. 1953. *A Value for n-Person Games*, pages 307–317. Princeton University Press.

Dan I Slobin. 2004. The many ways to search for a frog. *Relating events in narrative*, 2:219–257.

Andrew Spencer. 1991. *Morphological theory: An introduction to word structure in generative grammar*. Wiley-Blackwell.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Junyi Sun. 2012. Jieba. *Chinese word segmentation tool*.

Emiel van Miltenburg, Ákos Kádár, Ruud Koolen, and Emiel Krahmer. 2018a. DIDEC: The Dutch image description and eye-tracking corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3658–3669, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Emiel van Miltenburg, Ruud Koolen, and Emiel Krahmer. 2018b. Varying image description tasks: spoken versus written descriptions. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 88–100, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Serkan Çakmak. 2013. "var" ve "yok" sözcüklerinin morfolojik kimliği. *International Periodical For The Languages, Literature and History of Turkish or Turkic*, 8(4):463–471.

7 Supplemental Material

	Wikipedia	Caption
Arabic	11.60	4.65
Chinese	21.13	10.63
Farsi	24	7
German	13.43	9.47
Russian	15.43	4.27
Turkish	12.76	10.90

Table 4: Wikipedia sentences are on average longer, i.e. contain more tokens, than captions.

	Aspect	Arabic		Chinese		Farsi		German		Russian		Turkish	
		Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki	Capt.	Wiki
fastText	Atelic	0.84	-	0.12	-	0.67	-	**0.83**	-	**0.80**	-	**0.45**	-
	Telic	-	**0.67**	-	0.62	-	0.28	-	**0.79**	-	0.62	-	0.65
	State	0.10	0.66	0.53	**0.89**	0.46	0.59	**0.76**	0.71	0.27	**0.84**	0.42	0.53
mBERT	Atelic	**0.88**	-	**0.59**	-	**0.88**	-	0.48	-	0.78	-	0.44	-
	Telic	-	0.63	-	**0.79**	-	**0.69**	-	0.68	-	**0.84**	-	0.70
	State	**0.21**	**0.76**	**0.85**	0.41	**0.62**	**0.61**	0.66	**0.78**	0.17	0.50	**0.65**	**0.84**
ELMo	Atelic	0.85	-	0.00	-	0.45	-	0.79	-	0.17	-	0.38	-
	Telic	-	0.00	-	0.54	-	0.40	-	0.27	-	0.29	-	**0.82**
	State	0.10	0.74	0.81	0.00	0.35	0.46	0.00	0.55	**0.56**	0.55	0.48	0.00

Table 5: Zero-shot cross-lingual F1-scores per label across all languages with using fastText embeddings (top), multilingual BERT embeddings (middle) and ELMo embeddings (bottom).