Scalar Adjective Identification and Multilingual Ranking

Aina Garí Soler
Université Paris-Saclay
CNRS, LISN
91400, Orsay, France
aina.gari@limsi.fr

Marianna Apidianaki
Department of Digital Humanities
University of Helsinki
Helsinki, Finland
marianna.apidianaki@helsinki.fi

Abstract

The intensity relationship that holds between scalar adjectives (e.g., *nice* < *great* < *wonderful*) is highly relevant for natural language inference and common-sense reasoning. Previous research on scalar adjective ranking has focused on English, mainly due to the availability of datasets for evaluation. We introduce a new multilingual dataset in order to promote research on scalar adjectives in new languages. We perform a series of experiments and set performance baselines on this dataset, using monolingual and multilingual contextual language models. Additionally, we introduce a new binary classification task for English scalar adjective identification which examines the models' ability to distinguish scalar from relational adjectives. We probe contextualised representations and report baseline results for future comparison on this task.

1 Introduction

Scalar adjectives relate the entities they modify to specific positions on the evoked scale (e.g., GOODNESS, TEMPERATURE, SIZE): A *wonderful view* is nicer than a *good view*, and one would probably prefer a *delicious* to a *tasty meal*. But not all adjectives express intensity or degree. Relational adjectives are derived from nouns (e.g., *wood* → *wooden*, *chemistry* → *chemical*), have no antonyms and serve to classify nouns (e.g., *a wooden table*, *a chemical substance*) (McNally and Boleda, 2004). The distinction between scalar and relational adjectives is an important one. Identifying adjectives that express intensity can serve to assess the emotional tone of a given text, as opposed to words that mostly contribute to its descriptive content. Additionally, estimating the intensity of a scalar adjective is useful for textual entailment (*wonderful* $\models$ *good* but *good* $\not\models$ *wonderful*), product review analysis and recommendation systems, emotional chatbots and question answering (de Marneffe et al., 2010).

<table>
<tr><th></th><th>DEMELO</th></tr>
<tr><td>EN</td><td>dim &lt; gloomy < dark &lt; black</td></tr>
<tr><td>FR</td><td>terne &lt; sombre < foncé &lt; noir</td></tr>
<tr><td>ES</td><td>sombrío < tenebroso < oscuro &lt; negro</td></tr>
<tr><td>EL</td><td>αμυδρός || αχνός < μουντός < σκοτεινός< μαύρος</td></tr>
<tr><td></td><td>WILKINSON</td></tr>
<tr><td>EN</td><td>bad &lt; awful < terrible < horrible</td></tr>
<tr><td>FR</td><td>mauvais < affreux < terrible < horrible</td></tr>
<tr><td>ES</td><td>malo &lt; terrible < horrible < horroroso</td></tr>
<tr><td>EL</td><td>κακός &lt; απαίσιος < τρομερός < φρικτός</td></tr>
</table>

Table 1: Example translations from each dataset. "||" indicates adjectives at the same intensity level (ties).

Work on scalar adjectives has until now evolved around pre-compiled datasets (de Melo and Bansal, 2013; Taboada et al., 2011; Wilkinson and Oates, 2016; Cocos et al., 2018). Reliance on external resources has also restricted research to English, and has led to the prevalence of pattern-based and lexicon-based approaches. Recently, Garí Soler and Apidianaki (2020) showed that BERT representations (Devlin et al., 2019) encode intensity relationships between English scalar adjectives, paving the way for applying contextualised representations to intensity detection in other languages.[1]

In our work, we explicitly address the scalar adjective identification task, overlooked until now due to the focus on pre-compiled resources. We furthermore propose to extend scalar adjective ranking to new languages. We make available two new benchmark datasets for scalar adjective identification and multilingual ranking: (a) SCAL-REL, a balanced dataset of relational and scalar adjectives which can serve to probe model representations for scalar adjective identification; and (b) MULTI-SCALE, a scalar adjective dataset in French, Spanish and Greek. In order to test contextual models

[1] de Melo and Bansal (2013) discuss the possibility of a pattern-based multilingual approach which would require the translation of English patterns (e.g., "X but not Y") into other languages.

Proceedings of the 2021 Conference of the North American Chapter of the
Association for Computational Linguistics: Human Language Technologies, pages 4653–4660
June 6–11, 2021. ©2021 Association for Computational Linguistics

on these two tasks, the adjectives need to be seen in sentential context. We thus provide, alongside the datasets, sets of sentences that can be used to extract contextualised representations in order to promote model comparability. We conduct experiments and report results obtained with simple baselines and state-of-the-art monolingual and multilingual models on these new benchmarks, opening up avenues for research on sentiment analysis and emotion detection in different languages.[2]

		# unordered pairs	# adjectives
DEMELO	EN	548 (524)	339 (293)
	FR	590 (567)	350 (303)
	ES	448 (431)	313 (275)
	EL	557 (535)	342 (295)
WILKINSON	EN	61 (61)	59 (58)
	FR	67 (67)	61 (60)
	ES	59 (59)	58 (56)
	EL	68 (68)	61 (58)

Table 2: Composition of the translated datasets. In parentheses, we give the number of unique adjectives and pairs.

2 The Datasets

2.1 The MULTI-SCALE Dataset

We translate two English scalar adjective datasets into French, Spanish and Greek: DEMELO consists of 87 hand crafted half-scales[3] (de Melo and Bansal, 2013) and WILKINSON contains 12 full scales (Wilkinson and Oates, 2016). We use the partitioning of WILKINSON into 21 half-scales proposed by Cocos et al. (2018). In what follows, we use the term "scale" to refer to half-scales.

The two translators have (near-)native proficiency in each language. They were shown the adjectives in the context of a scale. This context narrows down the possible translations for polysemous adjectives to the ones that express the meaning described inside the scale. For example, the Spanish translations proposed for the adjective *hot* in the scales {*warm* < *hot*} and {*flavorful* < *zesty* < *hot* || *spicy*} are *caliente* and *picante*, respectively. Additionally, the translators were instructed to preserve the number of words in the original scales when possible. In some cases, however, they proposed alternative translations for English words, or none if an adequate translation could not be found. As a result, the translated datasets have a different number of words and ties. Table 1 shows examples of original English scales and their French, Spanish and Greek translations. Table 2 contains statistics on the composition of the translated datasets.

In order to test contextual models on the ranking task, we collect sentences containing the adjectives from OSCAR (Suárez et al., 2019), a multilingual corpus derived from CommonCrawl. French, Spanish and Greek are morphologically rich languages where adjectives need to agree with the noun they modify. In order to keep the method resource-light, we gather sentences that contain the adjectives in their unmarked form.

For each scale s, we randomly select ten sentences from OSCAR where adjectives from s occur. Then, we generate additional sentences through lexical substitution. Specifically, for every sentence (context) c that contains an adjective a_i from scale s, we replace a_i with $\forall$ $a_j \in s$ where $j = 1...|s|$ and $j \neq i$. This process results in a total of $|s|$ * 10 sentences per scale and ensures that $\forall$ $a \in s$ is seen in the same ten contexts. For English, we use the ukWaC-Random set of sentences compiled by Garí Soler and Apidianaki (2020) which contains sentences randomly collected from the ukWaC corpus (Baroni et al., 2009).

2.2 The SCAL-REL Dataset

SCAL-REL contains scalar adjectives from the DEMELO, WILKINSON and CROWD (Cocos et al., 2018) datasets (i.e. 79 additional half-scales compared to MULTI-SCALE). We use all unique scalar adjectives in the datasets (443 in total), and subsample the same number of relational adjectives, which are labelled with the pertainym relationship in WordNet (Fellbaum, 1998). There are 4,316 unique such adjectives in WordNet, including many rare or highly technical terms (e.g., *birefringent*, *anaphylactic*).[4] Scalar adjectives in our datasets are much more frequent than these relational adjectives; their average frequency in Google Ngrams (Brants and Franz, 2006) is 27M and 1.6M, respectively. We balance the relational adjectives set by frequency, by subsampling 222 frequent and 221 rare adjectives. We use the mean frequency of the

[2] Our code and data are available at https://github.com/ainagari/scalar_adjs.

[3] A full scale (e.g., {*hideous* > *ugly*, *pretty* < *beautiful* < *gorgeous*} can be split into two half scales which contain antonyms, often expressing different polarity {*hideous* > *ugly*} and {*pretty* < *beautiful* < *gorgeous*}.

[4] Note that the WordNet annotation does not cover all pertainyms in English (for example, frequent words such as *ironic* or *seasonal* are not marked with this relation).

4,316 relational adjectives in Google Ngrams as a threshold.[5] We propose a train/dev/test split of the SCAL-REL dataset (65/10/25%), observing a balance between the two classes (scalar and relational) in each set. To obtain contextualised representations, we collect for each relational adjective ten random sentences from ukWaC. For scalar adjectives, we use the ukWaC-Random set of sentences (cf. Section 2.1).

3 Multilingual Scalar Adjective Ranking

3.1 Methodology

Models We conduct experiments with state-of-the-art contextual language models and several baselines on the MULTI-SCALE dataset. We use the pre-trained `cased` and `uncased` multilingual BERT model (Devlin et al., 2019) and report results of the best variant for each language. We also report results obtained with four monolingual models: `bert-base-uncased` (Devlin et al., 2019), `flaubert_base_uncased` (Le et al., 2020), `bert-base-spanish-wwm-uncased` (Cañete et al., 2020), and `bert-base-greek-uncased-v1` (Koutsikakis et al., 2020). We compare to results obtained using fastText static embeddings in each language (Grave et al., 2018).

For a scale s, we feed the corresponding set of sentences to a model and extract the contextualised representations for $\forall\ a \in s$ from every layer. When an adjective is split into multiple BPE units, we average the representations of all wordpieces (we call this approach "WP") or all pieces but the last one ("WP-1"). The intuition behind excluding the last WP is that the ending of a word often corresponds to a suffix with morphological information.

The DIFFVEC method We apply the adjective ranking method proposed by Garí Soler and Apidianaki (2020) to our dataset, which relies on an intensity vector (called $\overrightarrow{dVec}$) built from BERT representations. The method yields state-of-the art results with very little data; this makes it easily adaptable to new languages. We build a sentence specific intensity representation ($\overrightarrow{dVec}$) by subtracting the vector of a mild intensity adjective, a_{mild} (e.g., *smart*), from that of a_{ext}, an extreme adjective on the same scale (e.g., *brilliant*) in the same context. We create a $dVec$ representation from every sentence available for these two reference adjectives, and average them to obtain the global $\overrightarrow{dVec}$ for that pair. Garí Soler and Apidianaki (2020) showed that a single positive adjective pair (DIFFVEC-1 $(+)$) is enough for obtaining highly competitive results in English. We apply this method to the other languages using the translations of a positive English (a_{mild}, a_{ext}) pair from the Crowd dataset: *perfect-good*.[6]

Additionally, we learn two dataset specific representations: one by averaging the $\overrightarrow{dVec}$'s of all (a_{ext}, a_{mild}) pairs in Wilkinson that do not appear in deMelo (DIFFVEC-WK), and another one from pairs in deMelo that are not in Wilkinson (DIFFVEC-DM). We rank adjectives in a scale by their cosine similarity to each $\overrightarrow{dVec}$: The higher the similarity, the more intense the adjective is.

Baselines We compare our results to a frequency and a polysemy baseline (FREQ and SENSE). These baselines rely on the assumption that low intensity words (e.g., *nice, old*) are more frequent and polysemous than their extreme counterparts (*e.g., awesome, ancient*). Extreme adjectives often limit the denotation of a noun to a smaller class of referents than mild intensity adjectives (Geurts, 2010). For example, an "awesome view" is more rare than a "nice view". This assumption has been confirmed for English in Garí Soler and Apidianaki (2020). FREQ orders words in a scale according to their frequency: Words with higher frequency have lower intensity. Given the strong correlation between word frequency and number of senses (Zipf, 1945), we also expect highly polysemous words (which are generally more frequent) to have lower intensity. This is captured by the SENSE baseline which orders the words according to their number of senses: Words with more senses have lower intensity.

Frequency is taken from Google Ngrams for English, and from OSCAR for the other three languages. The number of senses is retrieved from WordNet for English, and from BabelNet (Navigli and Ponzetto, 2012) for Spanish and French.[7] For adjectives that are not present in BabelNet, we use a default value which corresponds to the average number of senses for adjectives in the dataset (deMelo or Wilkinson) for which this information is available. We omit the SENSE baseline for

[5] Nine scalar adjectives from our datasets are also annotated as pertainyms in WordNet (e.g., *skinny, microscopic*) because they are denominal. We consider these adjectives to be scalar for our purposes since they clearly belong to intensity scales.

[6] FR: parfait-bon, ES: perfecto-bueno, EL: τέλειος-καλός.

[7] We omit Named Entities from BabelNet entries (e.g., names of TV shows or locations).

		EN			FR			ES			EL		
		Mono WP-1			**Mono WP-1**			**Mono WP-1**			**Mono WP-1**		
		P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
DM	DV-1 (+)	**.651**$_9$	**.435**$_9$	**.496**$_9$	**.610**$_3$	**.369**$_3$	**.396**$_3$	.658$_9$	.381$_9$	**.407**$_9$	.564$_2$	.238$_1$	.271$_2$
	DV-WK	.586$_6$	.267$_6$	.300$_6$	.515$_1$	.167$_1$	.166$_7$	**.670**$_7$	**.404**$_7$	**.407**$_7$	.589$_2$	.294$_2$	.325$_2$
WK	DV-1 (+)	.852$_1$	.705$_1$	.802$_1$	.612$_6$	.257$_6$	.215$_6$	**.814**$_7$	**.627**$_7$	**.803**$_9$	.618$_8$	.282$_8$	.256$_8$
	DV-DM	**.918**$_{10}$	**.836**$_{10}$	**.859**$_{10}$	.642$_7$	.322$_2$	.392$_2$	.780$_6$	.559$_6$	.684$_6$	**.750**$_{10}$	**.564**$_{10}$	**.586**$_{10}$
		Multi WP-1			**Multi WP**			**Multi WP**			**Multi (unc) WP**		
DM	DV-1 (+)	.609$_4$	.346$_4$	.389$_4$	.559$_7$	.260$_7$	.311$_7$	.614$_3$	.291$_3$	.268$_5$	.517$_9$	.139$_9$	.163$_9$
	DV-WK	.544$_3$	.208$_3$	.241$_4$	.517$_{10}$	.170$_{10}$	.179$_{10}$	.618$_{12}$	.301$_{12}$	.303$_{12}$	.539$_9$	.181$_9$	.207$_9$
WK	DV-1 (+)	.836$_6$	.672$_6$	.717$_6$	.672$_3$	.382$_3$	.380$_3$	.797$_3$	.593$_3$	.639$_3$	.662$_{10}$	.388$_9$	.423$_9$
	DV-DM	.836$_7$	.672$_7$	.766$_7$	**.701**$_6$	**.441**$_6$	**.476**$_2$	.695$_{10}$	.390$_{10}$	.511$_{10}$	.691$_5$	.447$_5$	.502$_5$
		Static models and baselines											
DM	DV-1 (+)	.637	.407	.458	.573	.288	.275	.656	.383	**.421**	.575	.266	.273
	DV-WK	.599	.330	.406	.454	.033	-.006	.616	.298	.315	.549	.205	.217
	FREQ	.575	.271	.283	.602	.346	.345	.585	.227	.239	**.596**	**.306**	**.334**
	SENSE	.493	.163	.165	.512	.229	.185	.516	.139	.151	-	-	-
WK	DV-1 (+)	.787	.574	.663	.582	.197	.152	.695	.390	.603	.706	.464	.566
	DV-DM	.852	.705	.783	.642	.325	.280	.712	.424	.547	.691	.447	.451
	FREQ	.754	.508	.517	.567	.167	.148	.576	.153	.382	.676	.417	.427
	SENSE	.721	.586	.575	.567	.255	.340	.644	.411	.456	-	-	-

Table 3: Results of the DIFFVEC (DV) method with monolingual (Mono) and multilingual (Multi) contextual models. Comparison to static embeddings and baselines per language. Subscripts denote the best layer. The best result obtained for each dataset in each language is indicated in boldface. For all languages but Greek, the multilingual model is cased.

Greek due to low coverage.[8]

3.2 Evaluation

We use evaluation metrics traditionally used for ranking evaluation (de Melo and Bansal, 2013; Cocos et al., 2018): Pairwise accuracy (P-ACC), Kendall's τ and Spearman's ρ. Results on this task are given in Table 3. Monolingual models perform consistently better than the multilingual model, except for French. We report the best wordpiece approach for each model: WP-1 works better with all monolingual models and the multilingual model for English. Using all wordpieces (WP) is a better choice for the multilingual model in other languages. We believe the lower performance of WP-1 in these settings to be due to the fact that the multilingual BPE vocabulary is mostly English-driven; this naturally results in highly arbitrary partitionings in these languages (e.g., ES: *fantástico* → fant-ástico; EL: γιγάντιος (*gigantic*)→γ-ι-γ-άν-τιος). Tokenisers of the monolingual models instead tend to split words in a way that more closely reflects the morphology of the language (e.g., ES: *fantástico* → fantás-tico; EL: γιγάντιος→γιγά-ντι-ος. Detailed results are found in Appendix A.

We observe that DIFFVEC-1 (+) yields comparable and sometimes better results than DIFFVEC-DM and DIFFVEC-WK, which are built from multiple pairs. This is important especially in the multilingual setting, since it shows that just one pair of adjectives is enough for obtaining good results in a new language. The best layer varies across models and configurations. The monolingual French and Greek models generally obtain best results in earlier layers. A similar behaviour is observed for the multilingual model for English to some extent, whereas for the other models performance improves in the upper half of the Transformer network (layers 6-12). This shows that the semantic information relevant for adjective ranking is not situated at the same level of the Transformer in different languages. We plan to investigate this finding further in future work. The lower results in French can be due to the higher amount of ties present in the datasets compared to other languages.[9] The baselines obtain competitive results showing that the underlying linguistic intuitions hold across languages. The best models beat the baselines in all configurations except for Greek on the DEMELO dataset, where FREQ and static embeddings obtain higher results. Overall, results are lower than those

[8] Only 47% of the Greek adjectives have a BabelNet entry, compared to 95.7% and 88.9% for Spanish and French. All English adjectives are present in WordNet.

[9] 58% of the French DEMELO scales contain a tie, compared to 45% in English.

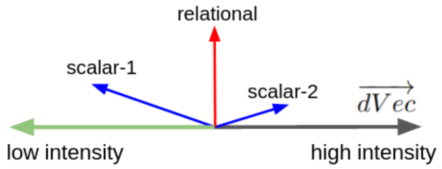

Figure 1: Illustration of two scalar adjectives that are close to $\overrightarrow{dVec}$ and to its opposite (which represents low intensity). The red vector describes a relational adjective that is perpendicular to $\overrightarrow{dVec}$.

reported for English, which shows that there is room for improvement in new languages.

4 Scalar Adjective Identification

For each English adjective in the SCAL-REL dataset, we generate a representation from the available ten sentences (cf. Section 2.2) using the `bert-base-uncased` model (with WP and WP-1). We experiment with a simple logistic regression classifier that uses the averaged representation for an adjective (ADJ-REP) as input and predicts whether it is scalar or relational. We also apply the DIFFVEC-1 (+) method to this task and measure how intense an adjective is by calculating its cosine with $\overrightarrow{dVec}$. The absolute value of the cosine indicates how clearly an adjective encodes the notion of intensity. In Figure 1, we show two scalar adjective vectors with negative and positive cosine similarity to $\overrightarrow{dVec}$, and another vector that is perpendicular to $\overrightarrow{dVec}$, i.e. describing a relational adjective for which the notion of intensity does not apply.[10] We train a logistic regression model to find a cosine threshold separating scalar from relational adjectives (DV-1 (+)). Finally, we also use as a feature the cosine similarity of the adjective representation to the vector of "*good*", which we consider as a prototypical scalar adjective (PROTO-SIM).

The best BERT layer is selected based on the accuracy obtained on the development set. We report accuracy on the test set. The baseline classifiers only use frequency (FREQ) and polysemy (SENSE) as features. We use these baselines on SCAL-REL because the WordNet pertainyms included in the dataset are rarer than the scalar adjectives. The intuition behind the SENSE baseline explained in Section 3.1 also applies here.

Method	Accuracy	
	WP	WP-1
ADJ-REP (BERT)	**0.946**$_9$	0.942$_9$
PROTO-SIM	0.888$_{11}$	0.902$_{10}$
DV-1 (+)	0.549$_2$	0.545$_2$
ADJ-REP (fastText)	0.929	
FREQ	0.669	
SENSE	0.714	

Table 4: Classification results on the SCAL-REL dataset.

Results on this task are given in Table 4. The classifier that relies on ADJ-REP BERT representations can distinguish the two types of adjectives with very high accuracy (0.946), closely followed by fastText embeddings (0.929). The DV-1 (+) method does not perform as well as the classifier based on ADJ-REP, which is not surprising since it relies on a single feature (the absolute value of the cosine between $\overrightarrow{dVec}$ and ADJ-REP). Comparing ADJ-REP to a typical scalar word (PROTO-SIM) yields better results than DV-1 (+). The SENSE and FREQ baselines can capture the distinction to some extent. Relational adjectives in our training set are less frequent and have fewer senses on average (2.59) than scalar adjectives (5.30). A closer look at the errors of the best model reveals that these concern tricky cases: One of the four misclassified scalar adjectives is derived from a noun (*microscopic*), whilst five out of eight wrongly classified relational adjectives can have a scalar interpretation (e.g., *sympathetic, imperative*). Overall, supervised models obtain very good results on this task. SCAL-REL will enable research on unsupervised methods that could be used in other languages.

5 Conclusion

We propose a new multilingual benchmark for scalar adjective ranking, and set performance baselines on it using monolingual and multilingual contextual language model representations. Our results show that adjective intensity information is present in the contextualised representations in the studied languages. We also propose a new classification task and a dataset that can serve as a benchmark to estimate the models' capability to identify scalar adjectives when relevant datasets are not available. We make our datasets and sentence contexts available to promote future research on scalar adjectives detection and analysis in different languages.

[10] To draw a parallel with gender debiasing, this value would reveal words' bias in the gender direction (Bolukbasi et al., 2016), regardless of the gender (male or female).

Acknowledgements

This work has been supported by the French National Research Agency under project ANR-16-CE33-0013. The work is also part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement № 771113). We thank the anonymous reviewers for their valuable suggestions.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Barcelona, Spain.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. In *LDC2006T13*, Philadelphia, Pennsylvania. Linguistic Data Consortium.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Anne Cocos, Skyler Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch. 2018. Learning Scalar Adjective Intensity from Paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows punta cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Bart Geurts. 2010. *Quantity implicatures*. Cambridge University Press.

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

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "Was It Good? It Was Provocative." Learning the Meaning of Scalar Adjectives". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics.

Louise McNally and Gemma Boleda. 2004. Relational adjectives as properties of kinds. Colloque de Syntaxe et Sémantique à Paris.

Gerard de Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Bryan Wilkinson and Tim Oates. 2016. A Gold Standard for Scalar Adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2669–2675, Portorož, Slovenia. European Language Resources Association (ELRA).

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *Journal of General Psychology*, 33(2):251–256.

A Comparison of Wordpiece Selection Methods

Table 3 of the main paper contains results of the DIFFVEC method with the best approach for selecting wordpieces (WPs) for each model. In Table 5, we present results obtained using the alternative approach for each model and language:

- for all monolingual models and the multilingual model for English, Table 5 contains results obtained with the WP approach;
- for the multilingual models in the other languages, we show results with WP-1.

The best approach was determined by comparing their average scores across the different methods. Some configurations improve, but they yield overall worse results per model, especially in Spanish. Differences between WP and WP-1 are generally more pronounced in the multilingual models than in the monolingual models.

		EN			FR			ES			EL		
		Mono WP			**Mono WP**			**Mono WP**			**Mono WP**		
		P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}	P-ACC	τ	ρ_{avg}
DM	DV-1 (+)	$\mathbf{.664}_{9}$	$\mathbf{.463}_{9}$	$\mathbf{.531}_{9}$	$\mathbf{.617}_{3}$	$\mathbf{.384}_{3}$	$\mathbf{.406}_{3}$	$\mathbf{.652}_{9}$	$\mathbf{.367}_{9}$	$\mathbf{.390}_{9}$	$.546_{8}$	$.201_{8}$	$.215_{8}$
	DV-WK	$.557_{9}$	$.246_{9}$	$.284_{6}$	$.517_{1}$	$.170_{1}$	$.140_{1}$	$.645_{10}$	$.353_{10}$	$.313_{10}$	$\mathbf{.557}_{2}$	$\mathbf{.226}_{2}$	$\mathbf{.240}_{2}$
WK	DV-1 (+)	$.852_{7}$	$.705_{7}$	$.766_{1}$	$.612_{7}$	$.262_{1}$	$.215_{6}$	$\mathbf{.763}_{8}$	$\mathbf{.525}_{8}$	$\mathbf{.755}_{6}$	$.632_{8}$	$.312_{8}$	$.256_{8}$
	DV-DM	$\mathbf{.918}_{6}$	$\mathbf{.836}_{6}$	$\mathbf{.839}_{6}$	$.627_{2}$	$.292_{2}$	$.392_{2}$	$.746_{6}$	$.492_{6}$	$.658_{6}$	$\mathbf{.779}_{1}1$	$\mathbf{.617}_{11}$	$\mathbf{.663}_{11}$
		Multi WP			**Multi WP-1**			**Multi WP-1**			**Multi (unc) WP-1**		
DM	DV-1 (+)	$.588_{4}$	$.301_{4}$	$.312_{4}$	$.549_{7}$	$.239_{7}$	$.276_{7}$	$.589_{3}$	$.229_{3}$	$.234_{1}$	$.524_{9}$	$.153_{9}$	$.171_{9}$
	DV-WK	$.516_{5}$	$.153_{11}$	$.198_{5}$	$.490_{2}$	$.113_{2}$	$.134_{7}$	$.603_{12}$	$.268_{12}$	$.287_{12}$	$.521_{6}$	$.146_{6}$	$.186_{6}$
WK	DV-1 (+)	$.820_{7}$	$.639_{7}$	$.667_{3}$	$.612_{3}$	$.262_{3}$	$.362_{3}$	$.746_{4}$	$.492_{4}$	$.608_{4}$	$.647_{9}$	$.358_{9}$	$.369_{9}$
	DV-DM	$.885_{7}$	$.770_{7}$	$.834_{7}$	$\mathbf{.687}_{7}$	$\mathbf{.412}_{7}$	$\mathbf{.435}_{3}$	$.661_{10}$	$.322_{10}$	$.447_{6}$	$.662_{6}$	$.388_{6}$	$.444_{6}$

Table 5: Results of DIFFVEC (DV) methods with contextualised representations derived from monolingual and multilingual models for each language, using an alternative approach to selecting wordpieces (WP, WP-1) than the one used for the results reported in Table 3. For all languages but Greek, the multilingual model is cased.