Automatic Estimation of Lexical Concreteness in 77 Languages

Bill Thompson (biltho@mpi.nl)
Gary Lupyan (glupyan@wisc.edu)
Language and Cognition Department, Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, Netherlands

Abstract

We estimate lexical Concreteness for millions of words across 77 languages. Using a simple regression framework, we combine vector-based models of lexical semantics with experimental norms of Concreteness in English and Dutch. By applying techniques to align vector-based semantics across distinct languages, we compute and release Concreteness estimates at scale in numerous languages for which experimental norms are not currently available. This paper lays out the technique and its efficacy. Although this is a difficult dataset to evaluate immediately, Concreteness estimates computed from English correlate with Dutch experimental norms at ρ = .75 in the vocabulary at large, increasing to ρ = .8 among Nouns. Our predictions also recapitulate attested relationships with word frequency. The approach we describe can be readily applied to numerous lexical measures beyond Concreteness.

Keywords: word2vec; Concreteness; multilingual; skipgram; norms

Introduction

What does a *chocolate cake* have in common with *deodorant*, and with *jumping*? Each of these words is, according to human raters, highly *Concrete* in the sense that these words represent phenomena that can be readily pointed to or enacted (Brysbaert, Warriner, & Kuperman, 2014). *Spirtuality*, on the other hand, like *inwardness* and *fun* are all abstract words the meanings of which are derived largely from language (Brysbaert, Warriner, & Kuperman, 2014; Brysbaert, Stevens, De Deyne, Voorspoels, & Storms, 2014). This dimension of Concreteness turns out to be one of the principal organizing dimensions of natural language vocabularies (Vankrunkelsven, Verheyen, De Deyne, & Storms, 2015).

Lexical norms represent normative judgments of the character of a word along an affective (e.g. emotional arousal), semantic (e.g. Concreteness), or social (e.g. usage frequency) dimension of interest. Analogous sets of lexical norms exist for a growing range of constructs: Valence, Arousal, and Dominance have been heavily studied (Warriner, Kuperman, & Brysbaert, 2013; Hollis, Westbury, & Lefsrud, 2017; Recchia & Louwerse, 2015a), for example, while more recent norm sets characterise constructs like Humour (Engelthaler & Hills, 2017), Iconicity (Winter, Perlman, Perry, & Lupyan, 2017), and Aversion (Thibodeau, 2016). Lexical norms support numerous research streams in the cognitive (e.g. (Larsen, Mercer, & Balota, 2006; Võ et al., 2009; Lodge & Taber, 2005; Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011)) and computational (e.g. (Staiano & Guerini, 2014; Esuli & Sebastiani, 2007)) sciences, but can be resource-intensive to collect experimentally (Hollis et al., 2017; Mandera, Keuleers, & Brysbaert, 2015), and are generally restricted to just one or several languages. Are there computational procedures that allow us to generalise experimental results beyond the relatively small vocabularies they often characterise? A number of researchers have begun pursuing this possibility, with promising results (Bestgen & Vincze, 2012; Hollis et al., 2017; Mandera et al., 2015; Recchia & Louwerse, 2015b, 2015a; Turney & Littman, 2003; Vankrunkelsven et al., 2015; Westbury et al., 2013; Bestgen, 2008; Feng, Cai, Crossley, & McNamara, 2011; Turney & Littman, 2002). Mandera et al. (2015) provides a recent overview and discussion of these techniques and their merits.

So far, this endeavor has largely focused on generalisation of experimental norms to larger vocabularies within a single language, such that the goal has been to incrementally increase predictive accuracy (and to evaluate alternative approaches to judging accuracy) on held-out experimental data with improved inferential techniques. These advances are reviewed in the next section. In this paper, we extend this enterprise to generalisation *across* languages. We show how it is possible to swap lexical norms between languages without the need to translate an entire vocabulary. This extension is made possible by the availability of semantic embeddings models in languages beyond English, and of techniques to align distinct embeddings spaces by translating a relatively small subset of their vocabularies. We combine these data and techniques with a variant of the most recent and powerful approach to norm-generalisation within languages. Doing so allows us to estimate lexical Concreteness in 77 languages. Although Concreteness is our focus, these relatively simple techniques are not limited to any particular lexical norm. As a result, we hope that the present work can function not only as a first-pass at estimating Concreteness cross-linguistically, but also as a demonstration of a technique that other researchers can apply to norm sets beyond Concreteness. All of our results and Python code to generate these and new predictions are availible at `github.com/billdthompson/cogsci-auto-norm`.

Previous Approaches to Norm Generalisation

In principle, estimating lexical norms for new vocabulary requires three ingredients: 1) a dataset of experimentally obtained lexical norms; 2) a machine-readable representation of relationships between words (this resource must cover at least some of the words that have been normed experimentally, plus a further set of words that have not been normed experimentally); and 3) an inferential procedure that facilitates norm prediction on the basis of 1) and 2). Thorough reviews of existing approaches to this task can be found in e.g. (Mandera et al., 2015; Hollis et al., 2017). Our method

also has close parallells in the computational linguistics literature, where norms of abstraction and imageability have been estimated via similar procedures and deployed in service of methaphor-detection accross langauges (Tsvetkov, Boytsov, Gershman, Nyberg, & Dyer, 2014)[1].

Corpora-based Approaches

Early approaches (Esuli & Sebastiani, 2007; Kim & Hovy, 2004) relied on linguistic resources such as WordNet (Miller, 1995), extrapolating norms to new words via synonymy relationships with already normed words. These approaches have been largely superseded by methods that make use of larger-scale text corpora (Turney & Littman, 2003; Bestgen, 2008; Bestgen & Vincze, 2012; Westbury et al., 2013; Recchia & Louwerse, 2015b), exploiting techniques such as Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), the Hyperspace Analogue to Language model (HAL) (Lund & Burgess, 1996), and the High Dimensional Explorer model (HiDEx) (Shaoul & Westbury, 2010, 2006) to estimate word similarities. Similarities derived from these models can then be used to bootstrap prediction of norms for new words, as a function of similarity to already-normed words (and sometimes of additional lexical properties too (Feng et al., 2011; Mandera et al., 2015)).

Recent Approaches

Mandera et al. (2015) provides perhaps the most systematic comparison between existing approaches. These authors consider four models from which relationships between words can be constructed: LSA (Deerwester et al., 1990), HAL (Lund & Burgess, 1996), topic modeling (Blei, Ng, & Jordan, 2003), and the recent Skipgram vector-embedding model (Mikolov, Chen, Corrado, & Dean, 2013). Two techniques for norm-generalisation were tested on each of these models: the *k*-nearest neighbors algorithm (Cover & Hart, 1967), and a random forests ensemble (Breiman, 2001). These results show that the best performing procedure is a combination of the Skipgram vector-embeddings model and a variant of the *k*-nearest neighbor algorithm.

Most recently, Hollis et al. (2017) extended the results from Mandera et al. (2015) by showing that the Skipgram based model performs even better when combined with a step-wise regression algorithm for inference. At the time of writing, Hollis et al. (2017)'s results represent the state of the art. By regressing experimental norms onto Skipgram-trained semantic vectors, and using the inferred regression coefficients to predict synthetic norms for newly observed vectors, Hollis and coleagues are able to make highly accurate predictions for Concreteness, arousal, and valence, among other norms (as established through cross-validation on held-out experimentally normed words). Among these norm sets, estimated norms of Concreteness in particular gain some of the highest correlations with held-out experimental data, on the order of $\rho = .86$. There has been some discussion of the bias imposed by this procedure. In particular, it has been noted that this procedure makes disproportionately more errors towards the extremes of human judgments (Hollis et al., 2017; Mandera et al., 2015). It has also been suggested that global correlation with human judgments alone, in papering over biases in estimation such as this, can profitably be accompanied by tests of agreement with independent lexical properties.

[1] We thank reviewer 3 for directing us to this literature.

Generalising Across Languages

The techniques outlined in the previous section show how it is possible to take a set of experimental norms, a computational models of semantics, and learn a mapping between these resources such that if the latter covers a larger set of words than the former, predictions can be made that exceed experimental vocabularies. In principle, the extension we outline amounts to observing that if the computational model of lexical semantics can include vocabulary from an *additional language*, then this same procedure can be used to estimate norms in the new language as well. Our approach in this paper should be understood as focusing on increasing the scale and applicability of these techniques, rather than on improving their accuracy.

We demonstrate our imputation procedure on Concreteness norms. Among lexical norm sets, Concreteness norms exist for an impresive number of words and have been colleceted in the same manner in two languages: English and Dutch. This facilitates cross-linguistic validation for at least this one case. Moreover, Concreteness is proving to be one of the principal organizing dimensions of vocabularies.

Methods

Training a Regression Model on Semantic Vectors

Suppose we have a set of empirically obtained point estimates (e.g. mean ratings) $\mathbf{y}^{\phi} = (y_1^{\phi}, \ldots, y_k^{\phi})$ on a dimension of interest ϕ (e.g. Concreteness) for a vocabulary $\mathbf{v} = (v_1, \ldots, v_k)$ of size k, such that y_i^{ϕ} is the experimental norm for vocabulary item v_i, for $i = 1, \ldots, k$. We obtain semantic embeddings $\vec{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ for each of these vocabulary items, such that $\mathbf{x_i}$ is the d-dimensional semantic vector for vocabulary item v_i. We train a simple regression model

$$y_i^{\phi} = \mathbf{x}_i \cdot \beta^{\phi} + \varepsilon_i \tag{1}$$

that relates $\mathbf{y}^{\phi}$ and $\vec{\mathbf{x}}$ through a noisy linear transformation (ε_i is isotropic Gaussian noise). After estimating $\hat{\beta}^{\phi} = (\hat{\beta}_1^{\phi}, \ldots, \hat{\beta}_d^{\phi})$, we can generate predictions, or synthetic norm estimates along the dimension ϕ, with

$$\mathbf{z}^{\phi} = \vec{\mathbf{x}} \cdot \hat{\beta}^{\phi} \ . \tag{2}$$

To extend these predictions to new words, we simply perform this transformations on a new set of word embeddings $\vec{\mathbf{x}}^*$.

Aligning Semantic Spaces

To make predictions about words in additional languages, we must be able to align embeddings vectors from multiple lan-

guages into the same semantic space, such that $\vec{\mathbf{x}}^*$ is multilingual. Typically, vector embeddings models only contain words in a single language. However, (Smith, Turban, Hamblin, & Hammerla, 2017) recently demonstrated that vector embeddings models for different languages can be *aligned* into a single underlying semantic space whenever it is possible to identify a subset of vocabulary items that represent *equivelant* points in space (translation equivalents). This general idea has recieved attention in the computational literature: see e.g. Ruder (2017).

Formally the procedure is as follows (see Smith et al. (2017) for full details). Let $\vec{\mathbf{x}}^{\mathbf{1}}$ be a set of semantic vectors in English, for example, and $\vec{\mathbf{x}}^{\mathbf{2}}$ be the set of semantic vectors for translation equivalents in Dutch (in the same order, so $\mathbf{x}_i^2$ is the vector for the Dutch translation of the English word with vector $\mathbf{x}_i^1$). If $\mathbf{m}$ is the matrix product of $\vec{\mathbf{x}}^{\mathbf{1}}$ and $\vec{\mathbf{x}}^{\mathbf{2}}$, and

$$\mathbf{u} \cdot \sigma \cdot \mathbf{v} = \mathbf{m} \tag{3}$$

represents a singular value decomposition of $\mathbf{m}$, then a transform to align the two semantic spaces can be obtained with $\mathbf{t} = \mathbf{u} \cdot \mathbf{v}$. Alignment is achieved through the product of any set of vectors in Dutch (or whichever language the transform has been trained on) and $\mathbf{t}$. This simple procedure allows us to align the semantics of any two languages for which we are able to obtain equivalent-dimensional word embeddings and a small set of translation equivalents. Following Smith et al. (2017), we take English to be the target language, and align other languages into this semantic space. After performing this transformation procedure on the word vectors for a new language, we can apply the generalisation procedure outlined in the previous section to word vectors in the new language (because $\vec{\mathbf{x}}^*$ now contains word vectors for words in the new language), even though the training set of vectors $\vec{\mathbf{x}}$ did not contain experimentally normed words in this language.

Embeddings Models

Our semantic models are based on the Facebook Artificial Intelligence Research (FAIR) release of embedding models. These models were each trained on Wikipedia data in the relevant language using the Skipgram technique. The quality of these models varies by language, reflecting in large part the size and variety of the available Wikipedia data for a given language. These are the resources we use as semantic models. We compute Concreteness estimates for a subset of these languages (77) based on the availability of comparable alignment transforms.

Alignment Transforms

Smith et al. (2017) recently released pre-computed alignment transforms $\mathbf{t}$ which position the vocabularies of a large subset of the FAIR languages into the same semantic space as the English model. These transforms were computed using Google-translate-obtained translations into English of the 10,000 most frequent lexical items in a given language. We tested alignment transforms that we computed ourselves in a handful of these languages using alternative, smaller sets of translation equivalents and found these to be equally effective in the main, but here we use the transforms released by Smith et al. (2017) throughout for consistency across languages, and because they are easily accessible.

Results

Concreteness in English and Dutch

Overall Agreement Applying our procedure[2] to Concreteness norms in English yields a correlation with English norms of $\rho = .86$, training on the full dataset of 33,286 (of 40,000) words for which we had both empirically collected data (Brysbaert, Warriner, & Kuperman, 2014) and we could obtain semantic vectors. This naturally aligns very closely with previously reported figures on related within-language prediction of English Concreteness (Hollis et al., 2017). Our predictions recapitulate the earlier finding that this procedure is biased to underestimate the extremes of human judgments (Mandera et al., 2015; Hollis et al., 2017). We also tested the same within-language procedure on Dutch experimental norms of Concreteness. We have not seen this case in the literature previously. We obtained Dutch vectors for 27772 (of 30,000) experimentally normed (Brysbaert, Stevens, et al., 2014) words. The correlation in this case is on the same order as predicting English: $\rho = .8$, and also demonstrates errors at the range extremes (this is inherent to the assumption of linearity in this particular technique). Most importantly, we trained the model on Concreteness norms in English, and applied the cross-linguistic generalisation procedure to predict Concreteness norms in *Dutch*. In this case, predictions correlate with Dutch norms at $\rho = .76$. Note also that in this case, these test data include held-out datapoints – datapoints not used to train the model - whereas in the within-language case, the correlations we report are best-possible-performance correlations (i.e. testing the model's predictions against the data on which it was trained). Given this distinction, and the fact that these cross-linguistic predictions were generated from experimental data collected among speakers of English, not Dutch, we view a correlation of $\rho = .76$ as constituting impressively high agreement.

Translation Equivelants & Lexical Properties In addition to these overall correlations, we explored the relationship between our predictions and experimental norms among a set of roughly 800 English-Dutch translation equivalents. We obtained these translations from the NORTHEURALEX dataset (Dellert & Jäger, 2017). This permits: 1) tests of language-specific prediction accuracy and 2) tests of language specific relationships with additional lexical properties. Figure 1 lays out the relationships among these variables. In this figure, each point represents a translation pair between English and Dutch, color coded for part of speech (Noun, Verb, or Ad-

[2]After applying the procedure, we rescale predictions into the legal range of the norms, i.e. 1 - 5 in the case of Concreteness.

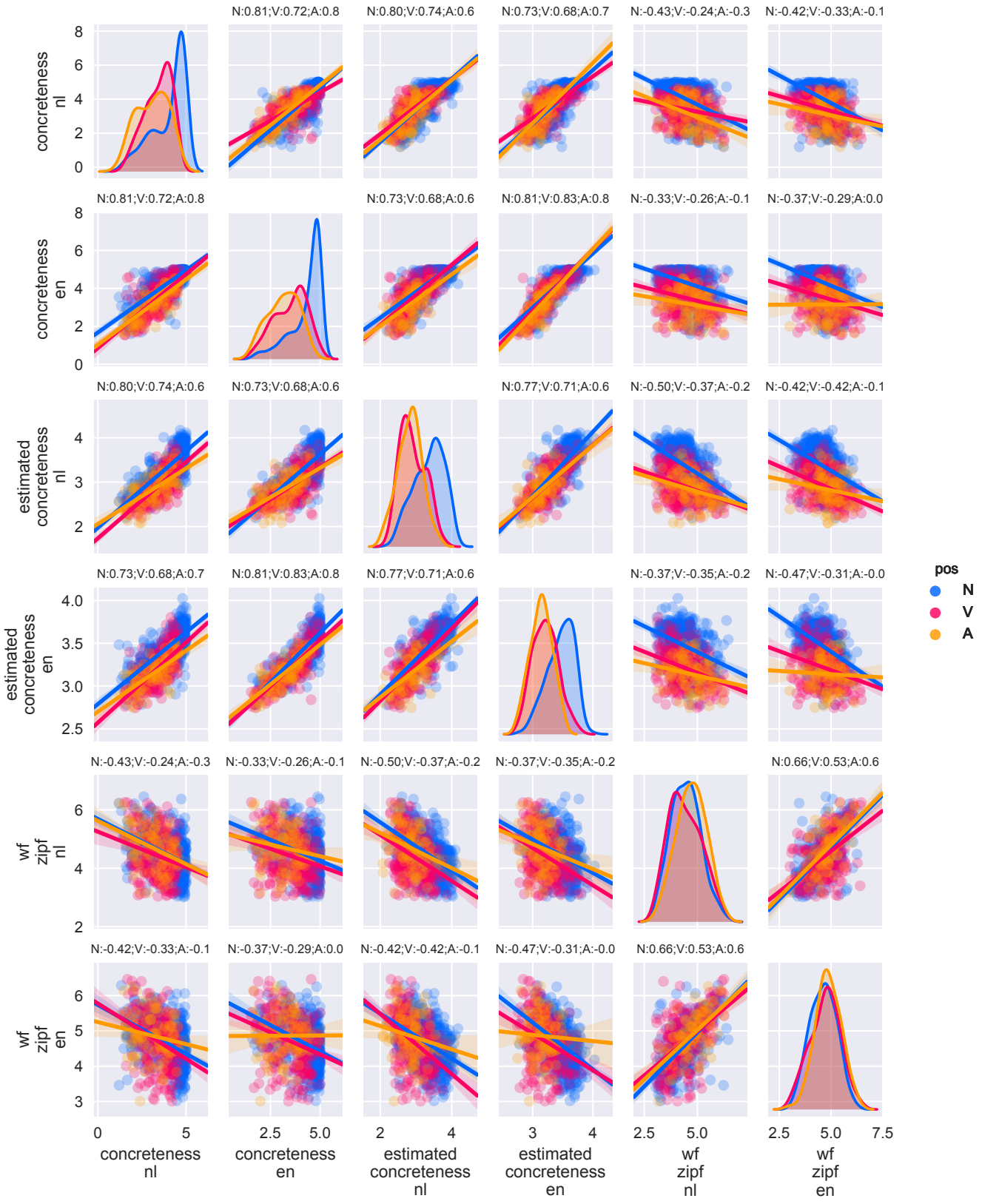

Figure 1: The relationships between experimental norms, estimated norms, and word frequency, across English and Dutch, among word pairs known to be translation equivalents, broken down by word class (N = Nouns, A = Adjectives, V = verbs). POS-specific Pearson correlation coefficients are given above each off-diagonal facet. Diagonal shows kernel density estimates.

Language	H	L
Catalan	mocador (4.90)	elogiar (2.18)
Czech	prkno (4.01)	velice (1.55)
Danish	spand (4.78)	aldrig (1.56)
Dutch	deksel (4.17)	nooit (2.03)
English	shoe (4.02)	urge (2.70)
Finnish	pussi (4.66)	vahva (1.58)
French	chaussure (4.60)	sembler (2.89)
German	backen (3.78)	niemals (1.91)
Hungarian	haj (4.68)	nevez (2.15)
Italian	coltello (4.51)	lodare (2.44)
Polish	torba (4.22)	bliski (1.84)
Portuguese	toalha (4.61)	alheio (2.17)
Romanian	os (3.22)	ceai (1.61)
Spanish	almohada (4.52)	sensato (2.29)
Swedish	ugn (4.14)	aldrig (2.07)
Turkish	kova (4.63)	zor (1.88)

Table 1: Words with high (H) and low (L) estimated Concreteness in 16 languages.

jective). The properties we explore are (log) word frequency (wf-log-zipf in figure 1) and part of speech (pos in figure 1). Although the figure contains numerous subtleties, here are a few key insights. First, Dutch Concreteness estimates (made on the basis of English norms) predict Dutch norms better than do estimates of English Concreteness (except among adjectives). The same is true in the other direction: estimates for English predict English norms better than do estimates for Dutch. These findings show the language-specificity of our prediction. Second, in both English and Dutch, Nouns tend to be more concrete than Verbs and Adjectives. This is true in our predicted ratings as well. Third, in both English and Dutch, more frequent Nouns and Verbs are less concrete. In Dutch, this also holds true for Adjectives, but the pattern is absent among English adjectives. Our predictions recapitulate these results as well.

Concreteness in 77 Languages

Although similar validation of imputed Concreteness values for all languages is, at this point, not possible, we release estimates of Concreteness in 77 languages. We expect the quality to vary by language. Since the model was trained on English norms, and the vector semantics aligned into English space, we expect the quality of the estimates to be better for languages more similar to English. Moreover, the vocabularies over which our estimates are computed are drawn from the Skipgram model vocabularies. Being scraped from the Internet, these vocabularies do contain errors. Nevertheless, we release these initial estimates in the hope that others with access to native speakers of these languages can establish some benchmarks, or explore the data computationally. Table 1 shows a sample of the most and least concrete words, as judged by our model in 16 languages.

Discussion & Conclusion

We showed how it is possible to swap lexical norms between languages. Our technique integrates existing techniques for estimating a relationship between distributional semantics models and experimental lexical norms, and using this relationship to estimate qualities of new words. Our principal contribution is to show how, when combined with methods for aligning vector spaces among distinct languages, and with contemporary embeddings models of distributional semantics in languages other than English, these techniques can be used to characterise vocabularies in new languages without the need for exhaustive translation. Though our validation capabilities are currently very limited, we showed that our predictions correlate with experimental observations in at least one other language, Dutch, and that we can recapitulate empirically observed relationships between Concreteness, word frequency, and word class, from our imputed values. We aim to test these predictions experimentally in future work, and hope that other researchers with access to speakers of these languages will be able to do the same. More generally, we hope that the technique laid out here may be of use to researchers beyond the example of Concreteness with which it has been illustrated. We see promise for this line of work in new methods for aligning vector spaces, and in improved computational procedures for distilling lexical norms into computational models.

References

Bestgen, Y. (2008). Building Affective Lexicons from Specific Corpora for Automatic Sentiment Analysis. In N. Calzolari et al. (Eds.), *Proceedings of lrec '08, 6th language resources and evaluation conference* (pp. 496–500). Marrakech, Morocco.: ELRA.

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. doi: 10.3758/s13428-012-0195-z

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

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

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica*, *150*, 80–84. doi: 10.1016/j.actpsy.2014.04.010

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. doi: 10.3758/s13428-013-0403-5

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. doi: 10.1109/TIT.1967.1053964

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic anal-

ysis. *Journal of the American society for information science*, *41*(6), 391.

Dellert, J., & Jäger, G. (2017). NorthEuraLex (version 0.9).

Engelthaler, T., & Hills, T. T. (2017). Humor norms for 4,997 English words. *Behavior Research Methods*, 1–9. doi: 10.3758/s13428-017-0930-6

Esuli, A., & Sebastiani, F. (2007). SentiWordNet: a high-coverage lexical resource for opinion mining. *Evaluation*, 1–26.

Feng, S., Cai, Z., Crossley, S. A., & McNamara, D. S. (2011). Simulating Human Ratings on Word Concreteness. In *Flairs conference.*

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, *70*(8), 1603–1619. doi: 10.1080/17470218.2016.1195417

Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on computational linguistics - coling '04* (pp. 1367–es). Morristown, NJ, USA: Association for Computational Linguistics. doi: 10.3115/1220355.1220555

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14–34. doi: 10.1037/a0021446

Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion*, *6*(1), 62–72. doi: 10.1037/1528-3542.6.1.62

Lodge, M., & Taber, C. S. (2005). The Automaticity of Affect for Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition Hypothesis. *Political Psychology*, *26*(3), 455–482. doi: 10.1111/j.1467-9221.2005.00426.x

Lund, K., & Burgess, C. (1996). Hyperspace analogue to language (HAL): A general model semantic representation. In *Brain and cognition* (Vol. 30, p. 5).

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, *68*(8), 1623–1642. doi: 10.1080/17470218.2014.988735

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Recchia, G., & Louwerse, M. M. (2015a). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598. doi: 10.1080/17470218.2014.941296

Recchia, G., & Louwerse, M. M. (2015b). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1584–1598. doi: 10.1080/17470218.2014.941296

Ruder, S. (2017). A survey of cross-lingual embedding models. *CoRR*, *abs/1706.0*.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190–195.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*(2), 393–413.

Smith, S. L., Turban, D. H. P., Hamblin, S., & Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint 1702.03859*.

Staiano, J., & Guerini, M. (2014). DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News.

Thibodeau, P. H. (2016). A Moist Crevice for Word Aversion: In Semantics Not Sounds. *PLOS ONE*, *11*(4), e0153686. doi: 10.1371/journal.pone.0153686

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., & Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 248–258).

Turney, P. D., & Littman, M. L. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *arXiv preprint cs/0212012*.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism. *ACM Transactions on Information Systems*, *21*(4), 315–346. doi: 10.1145/944012.944013

Vankrunkelsven, H., Verheyen, S., De Deyne, S., & Storms, G. (2015). Predicting lexical norms using a word association corpus. In *Proceedings of the 37th annual conference of the cognitive science society* (pp. 2463–2468). Cognitive Science Society.

Võ, M. L. H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538. doi: 10.3758/BRM.41.2.534

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. doi: 10.3758/s13428-012-0314-x

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: on emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in psychology*, *4*, 991. doi: 10.3389/fpsyg.2013.00991

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? *Interaction Studies*, *18*(3). doi: 10.1075/is.18.3.07win