

# Predicting Word Concreteness and Imagery

Jean Charbonnier  
Hochschule Hannover  
jean.charbonnier@hs-hannover.de

Christian Wartena  
Hochschule Hannover  
christian.wartena@hs-hannover.de

## Abstract

Concreteness of words has been studied extensively in psycholinguistic literature. A number of datasets has been created with average values for perceived concreteness of words. We show that we can train a regression model on these data, using word embeddings and morphological features. We evaluate the model on 7 publicly available datasets and show that concreteness and imagery values can be predicted with high accuracy. Furthermore, we analyse typical contexts of abstract and concrete words and review the potentials of concreteness prediction for image annotation.

## 1 Motivation

Concreteness and imagery of words has been studied for several decades in the field of psycholinguistics and psychology. Values for concreteness and imagery of words are obtained by instructing and asking experimentees to score words on a numeric scale for these aspects.

We assume that concrete nouns occur in other contexts than abstract nouns do and that nouns with a high imagery value occur together with other words than nouns with a low imagery. This is not as simple as it might sound, since most words can be used in different senses: e.g., nouns with high imagery might be accompanied by colour adjectives, but colours also fit perfectly with political parties and ideas. Nevertheless, we expect that there are differences in the distribution of abstract and concrete words and words with high and low imagery. If these differences indeed exist, and if concreteness and imagery are important aspects of the meaning of a word, we would expect that the characteristics of the context of concrete words are present in learned distributional representations of word meanings. This finally, can be verified quite easily and is exactly what we will test in the following: it should be possible to read off the concreteness and imagery of a word from its distributional representation.

If it is possible to predict the concreteness and imagery of a word from its distributional representation, this also has a very practical aspect: Retrieving these values from experiments is an expensive and time consuming task. If these values are needed for a psycholinguistic experiment or for some application it would be an advantage if we could compute them instead.

A practical application, that in fact was our initial motivation for this research, is the annotation of images (Charbonnier et al., 2018). We need to find terms in the caption (and surrounding text) of an image that describe that image. We expect that nouns with high concreteness and imagery values are much more likely to refer to concepts depicted in the image than abstract words do. Our basic intuition is illustrated by the image caption pair shown in Fig. 1. Here the noun *robot* in the caption is very concrete, while the other nouns (*systems*, *platform*, *research*, *development*) are much more abstract. The most concrete noun in this example describes quite well what is shown by the image.

The remainder of the paper is organised as follows: Section 2 gives an overview of common definitions of concreteness and imagery. In section 3 we review the most relevant literature on this topic. Section 4 gives an overview of the available datasets with human judgements on

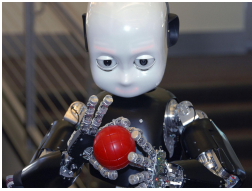
Image	Caption
	The iCub humanoid robot: an open-systems platform for research in cognitive development.
	Source
	Vernon, David, Michael Beetz, and Giulio Sandini. "Prospection in cognition: the case for joint episodic-procedural memory in cognitive robotics." <i>Frontiers in Robotics and AI</i> 2 (2015).

Figure 1: Typical image and caption from a scientific publication.

concreteness and imagery of words. In section 5 we present our method for predicting concreteness. The results are given in section 6. The source code for all experiments is available on GitHub<sup>1</sup>.

## 2 Concreteness and Imagery

Concreteness of words has received a lot of attention in psycholinguistic research. Here concreteness refers to the degree to which the concept denoted by a word refers to a perceptible entity (Brysbaert et al., 2014). Brysbaert et al. (2014) found that subjects largely rated the haptic and visual experiences, even if they were explicitly asked to take into account experiences involving any senses. Friendly et al. (1982) define concrete words as words that “refer to tangible objects, materials or persons which can be easily perceived with the senses”. They define imagery<sup>2</sup> as the ease with which a word arouses a mental image. Many studies found that there is a high correlation between concreteness and imagery (Friendly et al., 1982; Algarabel et al., 1988; Clark and Paivio, 2004).

It is assumed that concreteness influences learning, recognition memory and the speed of visual recognition, reading and spelling (Spreeen and Schulz, 1966). A recent overview of research in this area is e.g. given by Borghi et al. (2017).

## 3 Related Work

A few studies deal with the question whether values for concreteness can be predicted by machine learning techniques. Rabinovich et al. (2018) predict the concreteness of words indirectly by assigning a concreteness value to sentences in which a word occurs. The concreteness value of a sentence is based on the presence of seed words. The set of seed words is constructed by selecting words with derivational suffixes that are typical for highly abstract nouns. The correlation between manual assigned values from various subsets of the dataset from Brysbaert et al. (2014) (MT40k, see also Section 4.1) and the MRC database (see Section 4.2) and predicted values ranges from 0.66 to 0.74

Rothe et al. (2016) try to find low dimensional feature representations of words in which at least some dimensions correspond to interpretable properties of words. One of these dimensions is concreteness. For training and testing they use GoogleNews embeddings and two subsets of frequent words from the concreteness data by in MT40k. For their test set of 8,694 frequent words they found a moderate correlation with the human judgements and a value for Kendall’s  $\tau$  of 0,623.

Tanaka et al. (2013) use word concreteness to determine the reading difficulty of a text. Like we will do below, they train a regression model to predict concreteness values. As features they use a small number of manually constructed co-occurrence features, like co-occurrence with sense

<sup>1</sup><https://github.com/textmining-hsh/Concreteness>

<sup>2</sup>Most authors seem to use the term *imagery*, while others also use *imageability* and *visualness*. We will use *imagery* throughout this paper, even when describing datasets using one of the other terms.

verbs. For training and evaluation they use a subset of 3,455 nouns from the Medical Research Council Psycholinguistic Database (see Section 4 for details). Pearson’s correlation and Kendall’s  $\tau$  between the values from the database and their predictions are 0.675 and 0.502, respectively, for imagery, and 0.688 and 0.508, respectively, for concreteness.

Turney et al. (2011) proposed to use distributional vector representations as features to train a classifier that distinguishes abstract from concrete nouns. Turney constructed word embeddings for each word especially for this task. The recent work of Ljubešić et al. (2018) builds on this idea and tries to predict concreteness values instead of only considering two classes and uses standard word embeddings instead of training specialized word embeddings for the task. They found a Spearman correlation coefficient of 0.887 between the predicted concreteness values and the values from MT40k. The focus of their work is on transferring concreteness values from one to another language. In the current paper we will investigate this approach in more detail and evaluate on more datasets.

Hessel et al. (2018) use image captions to predict the likelihood of the occurrence of a word in the image and thus indirectly the concreteness or imagery of the word. In fact this exactly inverse method of our approach to annotating images: while Hessel et al. (2018) use likely descriptive terms to predict concreteness, we aim to find terms describing the image using concreteness.

## 4 Data

In order to support psycholinguistic research on differences of human processing of concrete and abstract words, for almost half a century researchers have collected concreteness values for words. The typical way to obtain these values is by averaging concreteness rates from several subjects in a controlled setting (Paivio et al., 1968). Recently, *Amazon Mechanical Turk* was used to get ratings for a large number of words (Brysbaert et al., 2014).

Despite the overlap between the available datasets we decided to evaluate our predictions on several, but not all collections, since all of them have their specific characteristics and have been used in other studies.

### 4.1 Datasets used for training and testing concreteness and imagery

We used four datasets for evaluation and one, the largest, for training. All datasets have values for concreteness, three of them also have values for imagery. To avoid confusion, we will treat the datasets with imagery and concreteness values as different datasets. Thus we have a total number of seven datasets. Table 1 gives an overview of the size of these datasets and their pairwise overlap. The table also shows for how many of the words in each of the datasets we find word embeddings in two common used resources of pretrained embeddings. Though we do not have enough imagery data to train a good model, we can, given the high correlation between imagery and concreteness values, also evaluate our concreteness model on imagery data. An overview of the correlation for words occurring in different data sets is provided in Table 2.

**MT40k** The dataset provided by Brysbaert et al. (2014) consist of 37,058 words and 2,896 two-word phrases rated by over 4,000 persons located in the USA using the online crowd sourcing tool *Amazon Mechanical Turk* (therefore we call this dataset MT40k). Each word was rated by at least 20 people. In the experiment 60,099 single words and 2,940 two-word expressions were used. Words that did not receive enough valid ratings got discarded. The remaining set of almost 40,000 English lemmas were known by at least 85% of the participants.

The results were validated with the concreteness values from the database of Coltheart (1981). For 3,935 words that are found in both collections, the Pearson correlation of the concreteness scores is  $r = 0.919$ . Unlike the other datasets below that use a scale from 1 to 7, MT40k ranges from 1 to 5. This scale was used because it was shown that 5 is the maximum number of

categories humans can work with reliably. The data is available as a CSV file<sup>3</sup> with all 60,099 words included. Words that did not receive enough valid ratings, are included in the file with a missing concreteness value.

**PYM<sub>C</sub>** The dataset created by Paivio et al. (1968) (PYM) consists of 925 nouns with ratings for concreteness, imagery and meaningfulness and was one of the first datasets for these values. Many other datasets are extension of this collection or used the same methodology to construct the dataset. In the following we denote words and concreteness values from this data set as PYM<sub>C</sub>. The data for PYM and CP are available as a CSV file<sup>4</sup>.

**PYM<sub>I</sub>** PYM<sub>I</sub> denotes the set all 925 nouns in PYM and their imagery values.

**CP<sub>A</sub>** Clark and Paivio (2004) collected and published various ratings and norms they could find for the 925 words from PYM. These ratings include also a previously unpublished set of imagery ratings (called IMG2 in their paper) that are different from the imagery values in PYM. We refer to this additional imagery ratings as CP<sub>A</sub>.

**CP<sub>E</sub>** Clark and Paivio (2004) also extended the word pool of Paivio et al. (1968) with more words, also including words with other part of speech than noun. The total size of the word pool is 2,311. For the new words ratings were collected in the same way as for the original words, including imagery ratings for 2,111 words. We refer to this extended dataset as CP<sub>E</sub>.

**TWP<sub>I</sub>** The Toronto Word Pool (TWP) by Friendly et al. (1982) consists of 1,080 common English words selected from the Thorndike-Lorge word count<sup>5</sup>. It includes not only nouns but also verbs, adjectives, adverbs and prepositions. Furthermore, the selected words all have a frequency of 20+ in Thorndike-Lorge and have a maximum of two syllables or eight letters. Only 20% of the words from PYM fulfil these restrictions. Hence the overlap between PYM and TWP is quite small. We refer to the TWP imagery ratings as TWP<sub>I</sub>. The experiment was done by 400 volunteering (160 male, 240 female) undergraduate psychology participants between 1977 and 1978. Every participant got one of 4 different lists with 270 words to rate. The values from TWP were extracted by using OCR and parsing from the scanned original paper.

**TWP<sub>C</sub>** TWP<sub>C</sub> denotes the concreteness values for all 1,080 words from TWP.

**Newcombe** Newcombe et al. (2012) constructed a dataset of 200 abstract and 200 concrete nouns, handpicked from TWP and PYM. The selected words have a concreteness and imagery rating of 5.0 or higher, whereas the abstract nouns are rated below 3.9. These data do not include any words with unclear concreteness and thus are intended to be used for experiment in which concrete and abstract words have to be contrasted. The words were extracted from the appendix of the paper with OCR and are manually corrected.

**Training Corpus** We constructed a set for training from the MT40K data set by removing all words from MT40K that also occur in either TWP, CP, PYM or Newcombe. In this way, we make sure that the words for which we predict imagery and concreteness, are never included in the training data. Furthermore we removed all two-phrase words and all words that are not part of the two resources for pretrained embeddings.

---

<sup>3</sup>[http://crr.ugent.be/papers/Concreteness\\_ratings\\_Brybaert\\_et\\_al\\_BRM.txt](http://crr.ugent.be/papers/Concreteness_ratings_Brybaert_et_al_BRM.txt)

<sup>4</sup><https://link.springer.com/article/10.3758/BF03195584#SupplementaryMaterial>

<sup>5</sup>The dataset TWP is available in the appendix of the original paper

Table 1: Size of datasets and the overlap with other data set.

	Size	$\cap$ Google	$\cap$ fastText	$\cap$ MT40k	$\cap$ TWP	$\cap$ PYM	$\cap$ CP <sub>E</sub>
MT40k	39954	33975	37058				
TWP	1080	1080	1080	1077			
PYM	925	921	925	877	167		
CP <sub>E</sub>	2111	2100	2111	1905	340	925	
Train	32783	31246	32783	32783	0	0	0

Table 2: Pearson correlation between the values for concreteness and imagery for the words in the intersection of two datasets.

	MT40k	TWP <sub>C</sub>	PYM <sub>C</sub>	TWP <sub>I</sub>	PYM <sub>I</sub>	CP <sub>A</sub>
TWP <sub>C</sub>	0.896					
PYM <sub>C</sub>	0.936	0.899				
TWP <sub>I</sub>	0.816	0.822	0.789			
PYM <sub>I</sub>	0.857	0.836	0.831	0.929		
CP <sub>A</sub>	0.717	0.731	0.596	0.897	0.803	
CP <sub>E</sub>	0.834	0.851	0.831	0.917	1.000	0.803

## 4.2 Further sources for concreteness

Besides the used datasets described above there are a number of further data sets that are aggregations of other datasets, very small, specialized or similar to newer data sets.

Spreen and Schulz (1966) determined concreteness ratings for 329 nouns. Gilhooly and Hay (1977) selected 205 five letter words with single-solution anagram with imagery and concreteness to analyse their effect on anagram solving. The Handbook of Semantic Word Norms (Toglia and Battig, 1978) gives concreteness values for 2,854 words. Gilhooly and Logie (1980) selected 1,944 nouns from Thorndike-Lorge word count and tried to have an even distribution over word length and frequency. Coltheart (1981) collected data from different publications to construct the Medical Research Council Psycholinguistic Database (MRCDB), a database with 98,538 words, 8,288 of which have values for imagery and concreteness originating from PYM, Toglia & Battig and Gilhooly & Logie. The Colorado Meaning Norms (Nickerson and Cartwright, 1984) contain 90 nouns from PYM and Toglia & Battig put into three concreteness groups (Low, Medium, High).

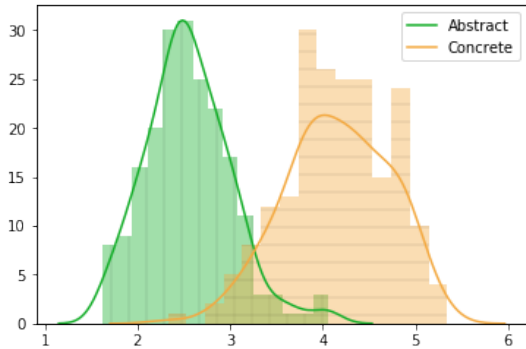
## 4.3 Embeddings

As distributional models for the words we use precomputed word embeddings from GoogleNews and fastText. The regression model will use the latent feature values from these embeddings to predict the concreteness values. GoogleNews embeddings were trained on a part of the GoogleNews dataset, which is about 100 billion words. The model contains 300-dimensional vectors for 3 million words and phrases (Mikolov et al., 2013).

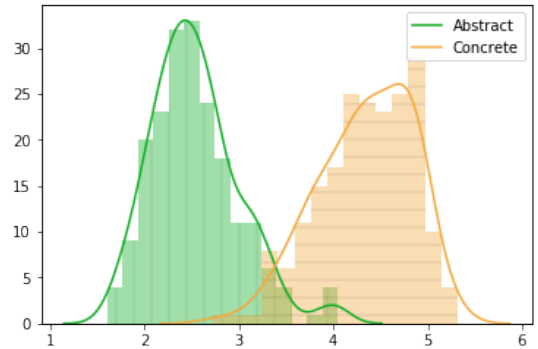
The fastText embeddings (Mikolov et al., 2018) are available in four versions. Two versions (with and without subword information) with 1 million word vectors are trained on Wikipedia 2017, UMBC webbase and statmt.org news with 16 billion tokens available. The other two version (also available with and without subword information) with 2 million word vectors trained on the Common Crawl with 600B tokens. In our experiments we used the version trained on Common Crawl without subword information, as it yields the best results.

## 5 Determining Word Concreteness

Given the availability of labelled data, the obvious way to predict concreteness is to train a regression model.



(a) Predictions by the classifier trained with GoogleNews word embeddings



(b) Predictions by the classifier trained with fastText word embeddings

Figure 2: Relative number of predicted concreteness values for abstract words (green) and concrete words (orange) from the Newcombe data.

## 5.1 Feature Selection

We identified three types of features, that could be useful for this task. In the first place, concrete words might occur in specific contexts, e.g. as object of *to see* or with adjectives like *green* or *wet*, etc. This fact was already used by Tanaka et al. (2013). Since the best context information for words currently available are word embeddings, we use word embeddings as features.

Furthermore, as noted by Rabinovich et al. (2018), certain suffixes can be important for determining concreteness. E.g. the suffix *-ness*, used to form a noun from an adjective, often refers to abstract concepts. Thus we take every possible suffix from within our training data with at least 1 character and at most 4 characters. We use the 200 most frequent suffixes as features.

Finally, the part of speech (POS) of a word might give a cue. Proper nouns, e.g., might more often refer to something concrete than verbs. Each word gets for each POS a value that is the relative frequency of all its lemmata for that POS found in WordNet.

## 5.2 Experimental setup

We trained a SVM to build a regression model. For the training we used  $\gamma = 0.01$ ,  $C = 1.0$  and an rbf kernel as parameters, found by grid search. We use the training corpus described above to train the regression model.

We evaluated the classifier using different sets of features with tenfold cross validation on the training data. Using all available features we evaluated the classifier on other datasets as well. In most datasets used for evaluation, there are a small number of words for which there are no pretrained word embeddings (see Table 1). For these words the SVM cannot predict a value. Hence, we will predict the value 3 (neutral, neither concrete nor abstract) for these words in the evaluation.

For the evaluation of all datasets with concreteness values we use Pearson’s  $r$  and Kendall’s  $\tau$  to measure the correlation between the true values and the model’s predictions. We took the output of our regression model as is, even if its prediction is outside of the target interval  $[1, 5]$  of our training data.

The dataset from Newcombe contains only binary data. Here we can order the words according to the predicted concreteness value and use the Area under the ROC Curve (AUC) as evaluation measure.

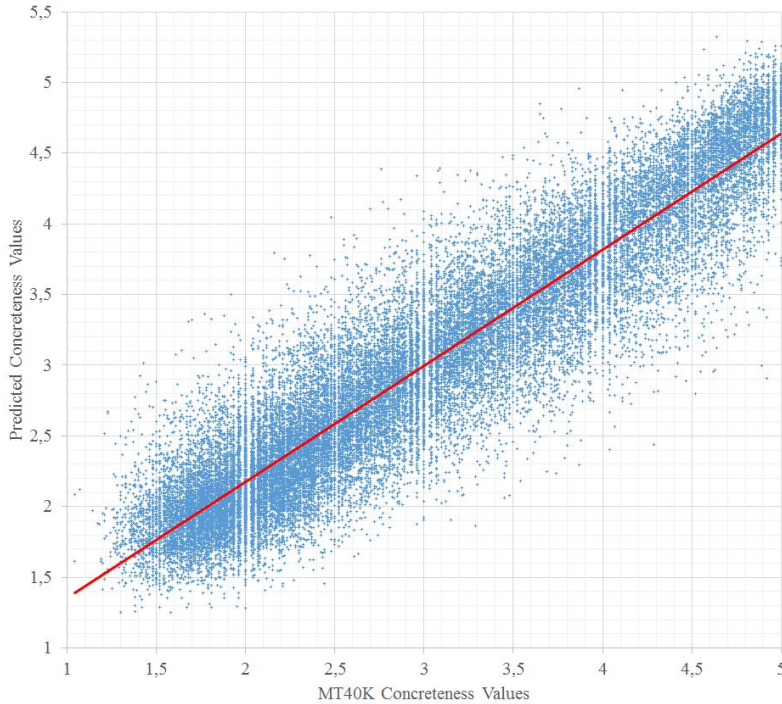


Figure 3: Original and predicted values (using 10 fold cross validation) for MT40K data using SVM with FastText word embeddings.

## 6 Results

Fig. 2 gives the distribution of predicted concreteness values for the concrete and abstract words from Newcombe and shows that the predicted values can distinguish quite well between concrete and abstract words. The AUC for this binary classification is 0.990 using fastText embeddings, POS and suffixes and 0.981 using GoogNews embeddings, POS and suffixes.

The correlation between the original and predicted MT40k concreteness values (aggregated from cross validation) is visualized in Fig. 3, and clearly shows the strong linear correlation. The corresponding correlation strength is given in Table 3. We see that the model trained with fastText gives consistently higher correlations than using GoogleNews embeddings. Adding the suffix or POS helps increasing the correlation for GoogleNews. Combining suffix and POS increases the performance also slightly, even for the already well performing fastText embeddings. Using only GoogleNews embeddings for cross validation on MT40k data, the (average) value for Kendall’s  $\tau$  is only 0.652, which is still in the same order of magnitude as the correlation found by Rothe et al. (2016) using the same features and a subset of frequent words from the MT40k data. Furthermore we see that the correlations found are much higher than those found by Tanaka et al. (2013) and Rabinovich et al. (2018). The result is comparable and just slightly higher than the correlation found by Ljubešić et al. (2018), who found a Spearman coefficient of 0.887. The Spearman coefficient for our best feature combination using fastText vectors is 0.900 with 10 fold cross validation.

The correlation of the predicted concreteness with the data from the other data sets is given in Table 4. Table 5 gives the correlation with the imagery datasets. Since the model was trained with concreteness values the smaller correlation for imagery scores is as expected. In fact, we see that the correlation values we found are consistently slightly below the correlation values for the overlapping parts of each dataset with MT40k, which shows that we are very close to the highest reasonably possible result. Note, that we excluded all words in the intersection of the datasets from the training data.

Finally, Table 6 gives some examples of predicted and original values for some abstract and

Table 3: Pearson ( $r$ ) and Kendall ( $\tau$ ) correlation results using our training corpus with cross-validation and different features.

Embedding	fastText		GoogleNews	
Correlation	$r$	$\tau$	$r$	$\tau$
pos + suffix	0.604	0.428	(0.604)	(0.428)
emb	0.905	0.711	0.856	0.652
emb + suffix	0.908	0.716	0.872	0.671
emb + pos	0.908	0.717	0.870	0.671
emb + pos + suffix	0.911	0.721	0.879	0.680

Table 4: Pearson ( $r$ ) and Kendall ( $\tau$ ) correlation between our concreteness estimations on the concreteness values of the TWP and PYM datasets using fastText and GoogleNews embeddings.

Embedding	fastText		GoogleNews	
Correlation	$r$	$\tau$	$r$	$\tau$
TWP <sub>C</sub>	0.881	0.698	0.852	0.656
PYM <sub>C</sub>	0.902	0.741	0.877	0.703

concrete words from the MT40k dataset in order to get an impression of involved words and values. We could not detect any pattern in the words for which the predictions differ a lot from the experimental values. We expected that the predictions might make many mistakes for those words where the experimentees disagreed a lot. In the MT40k the variance of all concreteness values is given, so this can be checked easily. We found that there is no correlation (Pearson correlation coefficient is 0.132) between the variance in the original data and the prediction error.

Table 5: Pearson ( $r$ ) and Kendall ( $\tau$ ) correlation between our imagery estimations on the imagery values of the TWP, PYM and CP datasets using fastText and GoogleNews embeddings.

Embedding	fastText		GoogleNews	
Correlation	$r$	$\tau$	$r$	$\tau$
TWP <sub>I</sub>	0.774	0.559	0.731	0.514
PYM <sub>I</sub>	0.813	0.618	0.770	0.568
CP <sub>A</sub>	0.676	0.499	0.619	0.453
CP <sub>E</sub>	0.796	0.569	0.745	0.521

## 7 Discussion and Future Work

We have shown that concreteness of words as perceived by a subject of a rating experiment can be predicted on the base of word embeddings. Besides contextual information, morphological cues turn out to help somewhat.

Word embeddings essentially encode information about the contexts a word appears in. Thus we can conclude that concrete words appear in different contexts than abstract words. Tanaka et al. (2013) e.g. assume that concrete words occur often in the context of sense verbs. In order to get an impression of the words that are typical for the context of abstract and concrete words we computed the concreteness values for a random selection of 5,000 words from ukWaC (Ferraresi et al., 2008) and selected the 200 most abstract and 200 most concrete words. For each of these 400 words we computed the positive pointwise mutual information (ppmi) with a set of 17,400 mid-frequency words for co-occurrence within a window of 2 words. For each word of this set of 17,400 words we compute the average ppmi with the abstract and concrete words, respectively. The words with high average ppmi values for concrete or abstract words are typical for the context of these 200 words. The words with the highest ppmi value are given in Table 7. As one can see, we found mainly material properties for very concrete words. For abstract words we



Table 6: Overview of high, medium and low concreteness for words from MT40K with their original and our predicted values.

rank	word	MT40k	predicted
1	watermelon	4.89	5.3246
2	hamburger	5.00	5.3123
3	postbox	4.54	5.2518
4	surfboard	4.57	5.2468
5	typewriter	4.88	5.2311
17711	magnetically	2.96	2.7366
17712	evenness	2.43	2.7364
17713	undrafted	2.63	2.7364
17714	distorted	2.57	2.7363
17715	amusement	2.07	2.7361
32779	inconceivable	1.38	1.2549
32780	irrelative	1.81	1.2532
32781	transcendental	1.48	1.2449
32782	notwithstanding	1.38	1.2225
32783	behooves	1.58	1.1129

Table 7: 30 words with highest pointwise mutual information in ukWaC with prototype of abstract and concrete words, resp.

Concrete			Abstract		
stuffed	plastic	dried	notions	purely	notion
wooden	lined	underneath	conceptions	theories	reasoning
giant	topped	coated	interpretation	manner	theory
black	leather	shaped	manifestations	concepts	rationality
underside	coloured	rubber	understanding	profound	nature
metal	bamboo	glass	rational	philosophical	expression
homemade	blue	washed	conception	analysis	utterly
bowl	red	mounted	discourses	manifestation	linguistic
decorated	yellow	steel	significance	aspects	aesthetic
white	bag	powder	expressions	psychological	discourse

found words such as *philosophical*, *conception*, *linguistic*, *discourse* and *theories*.

One of the goals of our project is to find good keywords for images from scientific publications. These images often have very long captions (see Sohmen et al., 2018). Thus, the captions and eventually the sentences explicitly referring to an image usually will provide enough text for extracting words describing the image (Josi et al., 2018).

To get a first impression of the potential of concreteness for finding words describing scientific images we consider the image that was also used by Josi et al. (2018), here shown in Fig. 4. Initially, 53 words and phrases (noun phrases that are titles of Wikipedia articles) were selected from the caption and referring context. Table 8 shows 10 terms with the highest idf values (in the complete collection of 2,9 million image captions) and the highest predicted concreteness values, resp. For two word phrases we used the maximum of the (predicted) concreteness values of the parts as the concreteness of the phrase. The idf values were computed directly for the phrases. In this example we clearly see the different aspects of both weighting schemes: idf favours specific terms that do not describe the image, like *Griffith university* and *Queensland*. Most words selected by high concreteness, describe quite well what can be seen on the image (except the most concrete word, *wood*), but are not specific enough: an arm and a rib are clearly present in the image, *deep fascia* is nevertheless a more adequate key word. Thus we expect that we will need to combine concreteness with other relevance measures for this application.

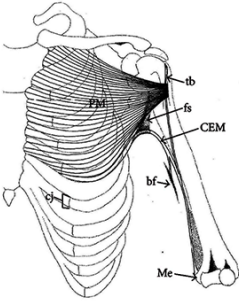
Image	Caption
	Schematic drawing of the left thorax and upper limb, demonstrating the chondroepitrochlearis muscle (CEM) inserting into the deep brachial fascia (bf) and the fibrous band (tuberoepicondylar band, tb) (PM: pectoralis major; fs: fascial sling; cj: costochondral junction; and Me: medial epicondyle).
	Source
	Sujeewa P. W. Palagama, Raymond A. Tedman, Matthew J. Barton, and Mark R. Forwood, "Bilateral Chondroepitrochlearis Muscle: Case Report, Phylogenetic Analysis, and Clinical Significance," <i>Anatomy Research International</i> , vol. 2016, Article ID 5402081, 2016.

Figure 4: Image and caption from a scientific publication used in Josi et al. (2018) to illustrate keyword extraction from image captions.

## Acknowledgements

This research was part of the NOA project and funded by the DFG under grant no. 315976924. We would like to thank the anonymous reviewers for their valuable feedback.

Table 8: Ranking of potential keywords selected from the caption and referring context of the images shown in Fig. 4 based on idf concreteness.

(a) Terms ranked by idf (in a collection of 2,9 Million image captions).

rank	term	idf
1	axillary fascia	20.0
2	griffith university	18.1
3	brachial fascia	17.5
4	quartus	15.7
5	medical literature	14.4
6	common name	13.9
7	deep fascia	13.9
8	epicodyle	12.7
9	joint capsule	12.4
10	queensland	12.2

(b) Terms with predicted concreteness and concreteness values from MT40k.

rank	term	pred. concr.	MT40k
1	wood	4.93	4.85
2	arm	4.80	4.96
3	biceps	4.68	4.93
4	rib	4.65	4.90
5	cartilage	4.43	4.71
6	thorax	4.43	4.56
7	tendon	4.40	4.47
8	cadaver	4.39	4.48
9	joint capsule	4.32	4.52
10	septum	4.26	4.48

## References

- Algarabel, S., J. C. Ruiz, and J. Sanmartin (1988). The University of Valencia’s computerized word pool. *Behavior Research Methods, Instruments, & Computers* 20(4), 398–403.
- Borghi, A. M., F. Binkofski, C. Castelfranchi, F. Cimatti, C. Scorolli, and L. Tummolini (2017). The challenge of abstract concepts. *Psychological Bulletin* 143(3), 263.
- Brysbaert, M., A. B. Warriner, and V. Kuperman (2014, September). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3), 904–911.
- Charbonnier, J., L. Sohmen, J. Rothman, B. Rohden, and C. Wartena (2018). Noa: A search engine for reusable scientific images beyond the life sciences. In G. Pasi, B. Piwowarski,

- L. Azzopardi, and A. Hanbury (Eds.), *Advances in Information Retrieval*, Cham, pp. 797–800. Springer International Publishing.
- Clark, J. M. and A. Paivio (2004, Aug). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers* 36(3), 371–383.
- Coltheart, M. (1981, November). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A* 33(4), 497–505.
- Ferraresi, A., E. Zanchetta, M. Baroni, and S. Bernardini (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pp. 47–54.
- Friendly, M., P. E. Franklin, D. Hoffman, and D. C. Rubin (1982, September). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation* 14(4), 375–399.
- Gilhooly, K. J. and D. Hay (1977, January). Imagery, concreteness, age-of-acquisition, familiarity, and meaningfulness values for 205 five-letter words having single-solution anagrams. *Behavior Research Methods & Instrumentation* 9(1), 12–17.
- Gilhooly, K. J. and R. H. Logie (1980, July). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation* 12(4), 395–427.
- Hessel, J., D. Mimno, and L. Lee (2018). Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2194–2205. Association for Computational Linguistics.
- Josi, F., C. Wartena, and J. Charbonnier (2018). Text-based annotation of scientific images using wikimedia categories. In M. Elloumi, M. Granitzer, A. Hameurlain, C. Seifert, B. Stein, A. M. Tjoa, and R. Wagner (Eds.), *Database and Expert Systems Applications*, Cham, pp. 243–253. Springer International Publishing.
- Ljubešić, N., D. Fišer, and A. Peti-Stantić (2018, July). Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of The Third Workshop on Representation Learning for NLP*, Melbourne, Australia, pp. 217–222. Association for Computational Linguistics.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mikolov, T., E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.
- Newcombe, P., C. Campbell, P. Siakaluk, and P. Pexman (2012). Effects of Emotional and Sensorimotor Knowledge in Semantic Processing of Concrete and Abstract Nouns. *Frontiers in Human Neuroscience* 6, 275.
- Nickerson, C. A. and D. S. Cartwright (1984, July). The University Of Colorado Meaning Norms. *Behavior Research Methods, Instruments, & Computers* 16(4), 355–382.
- Paivio, A., J. C. Yuille, and S. A. Madigan (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology* 76(1, Pt.2), 1–25.

- Rabinovich, E., B. Sznajder, A. Spector, I. Shnayderman, R. Aharonov, D. Konopnicki, and N. Slonim (2018, September). Learning Concept Abstractness Using Weak Supervision. *ArXiv e-prints*.
- Rothe, S., S. Ebert, and H. Schütze (2016). Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 767–777. Association for Computational Linguistics.
- Sohmen, L., J. Charbonnier, I. Blümel, C. Wartena, and L. Heller (2018). Figures in scientific open access publications. In E. Méndez, F. Crestani, C. Ribeiro, G. David, and J. C. Lopes (Eds.), *Digital Libraries for Open Knowledge*, Cham, pp. 220–226. Springer International Publishing.
- Spreeen, O. and R. W. Schulz (1966). Parameters of abstraction, meaningfulness, and pronounciability for 329 nouns. *Journal of Verbal Learning & Verbal Behavior* 5(5), 459–468.
- Tanaka, S., A. Jatowt, M. P. Kato, and K. Tanaka (2013). Estimating content concreteness for finding comprehensible documents. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, New York, NY, USA, pp. 475–484. ACM.
- Toglia, M. P. and W. F. Battig (1978). *Handbook of Semantic Word Norms*. Oxford, England: Lawrence Erlbaum.
- Turney, P. D., Y. Neuman, D. Assaf, and Y. Cohen (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, Stroudsburg, PA, USA, pp. 680–690. Association for Computational Linguistics.