Improving Verb Metaphor Detection by Propagating Abstractness to Words, Phrases and Individual Senses

Maximilian Köper and **Sabine Schulte im Walde**
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Germany
{maximilian.koeper,schulte}@ims.uni-stuttgart.de

Abstract

Abstract words refer to things that can not be seen, heard, felt, smelled, or tasted as opposed to concrete words. Among other applications, the degree of abstractness has been shown to be a useful information for metaphor detection. Our contribution to this topic are as follows: i) we compare supervised techniques to learn and extend abstractness ratings for huge vocabularies ii) we learn and investigate norms for multi-word units by propagating abstractness to verb-noun pairs which lead to better metaphor detection, iii) we overcome the limitation of learning a single rating per word and show that multi-sense abstractness ratings are potentially useful for metaphor detection. Finally, with this paper we publish automatically created abstractness norms for 3 million English words and multi-words as well as automatically created sense-specific abstractness ratings.

1 Introduction

The standard approach to studying abstractness is to place words on a scale ranging between abstractness and concreteness. Alternately, abstractness can also be given a taxonomic definition in which the abstractness of a word is determined by the number of subordinate words (Kammann and Streeter, 1971; Dunn, 2015).

In psycholinguistics abstractness is commonly used for concept classification (Barsalou and Wiemer-Hastings, 2005; Hill et al., 2014; Vigliocco et al., 2014). In computational work, abstractness has become an established information for the task of automatic detection of metaphorical language. So far metaphor detection has been carried out using a variety of features including selectional preferences (Martin, 1996; Shutova and Teufel, 2010; Shutova et al., 2010; Haagsma and Bjerva, 2016), word-level semantic similarity (Li and Sporleder, 2009; Li and Sporleder, 2010), topic models (Heintz et al., 2013), word embeddings (Dinh and Gurevych, 2016) and visual information (Shutova et al., 2016).

The underlying motivation of using abstractness in metaphor detection goes back to Lakoff and Johnson (1980), who argue that metaphor is a method for transferring knowledge from a concrete domain to an abstract domain. Abstractness was already applied successfully for the detection of metaphors across a variety of languages (Turney et al., 2011; Dunn, 2013; Tsvetkov et al., 2014; Beigman Klebanov et al., 2015; Köper and Schulte im Walde, 2016b).

The abstractness information itself is typically taken from a dictionary, created either by manual annotation or by extending manually collected ratings with the help of supervised learning techniques that rely on word representations. While potentially less reliable, automatically created norm-based abstractness ratings can easily cover huge dictionaries. Although some methods have been used to learn abstractness, literature lacks a comparison of these learning techniques.

We compare and evaluate different learning techniques. In addition we show and investigate the usefulness of extending abstractness ratings to phrases as well as individual word senses. We extrinsically evaluate these techniques on two verb metaphor detection tasks: (i) a type-based setting that makes use of phrase ratings, (ii) a token-based classification for multi-sense abstractness norms. Both settings benefit from our approach.

2 Experiments

2.1 Propagating Abstractness: A Comparison of Approaches & Ressources

2.1.1 Comparison of Approaches

Turney et al. (2011) first aproached to automatically create abstractness norms for 114 501 words, relying on manual ratings based on the MRC Psycholinguistic Database (Coltheart, 1981). The underlying algorithm (Turney and Littman, 2003) requires vector representation and annotated training samples of words. The algorithm itself performs a greedy forward search over the vocabulary to learn so-called paradigm words. Once paradigm words for both classes (abstract & concrete) are learned, a rating can be assigned to every word by comparing its vector representation against the vector representations of the paradigm words.

Köper and Schulte im Walde (2016a) used the same algorithm for a large collection of German lemmas, and in the same way additional created ratings for multiple norms including valency, arousal and imageability.

A different method that has been used to extend abstractness norms based on low-dimensional word embeddings and a Linear Regression classifier (Tsvetkov et al., 2013; Tsvetkov et al., 2014).

We compare approaches across different publicly available vector representations[1], to study potential differences across vector dimensionality we compare vectors between 50 and 300 dimensions. The Glove vectors (Pennington et al., 2014) have been trained on 6billion tokens of Wikipedia plus Gigaword (V=400K), while the word2vec cbow model (Mikolov et al., 2013) was trained on a Google internal news corpus with 100billion tokens (V=3million). For training and testing we relied on the ratings from Brysbaert et al. (2014), Dividing the ratings into 20% test (7 990) and 80% training (31 964) for tuning hyper parameters we took 1 000 ratings from the training data. We kept the ratio between word classes. Evaluation is done by comparing the new created ratings against the test (gold) ratings using Spearman's rank-order correlation. We first reimplemented the algorithm from Turney and Littman (2003) (`T&L 03`). Inspired by recent findings of Gupta et al. (2015) we apply the hypothesis that distributional vectors implicitly encode attributes such as abstractness and directly feed the vector representation of a word into a classifier, either by using linear regression (`L-Reg`), a regression forest (`Reg-F`) or a fully connected feed forward neural network with up to two hidden layers (`NN`).[2]

	T&L 03	L-Reg.	Reg-F.	NN
Glove50	.76	.76	.78	**.79**
Glove100	.80	.79	.79	**.85**
Glove200	.78	.78	.76	**.84**
Glove300	.76	.78	.74	**.85**
W2V300	.83	.84	.79	**.90**

Table 1: Spearman's ρ for the test ratings. Comparing representations and regression methods.

Table 1 shows clearly that we can learn abstractness ratings with a very high correlation on the test data using the word representations from Google (W2V300) together with a neural network for regression (ρ=.90). The `NN` method significantly outperforms all other methods, using Steiger (1980)'s test ($p < 0.001$).

2.1.2 Comparison of Ressources

Based on the comparison of methods in the previous section we propagated abstractness ratings to the entire vocabulary of the W2V300 dataset (3million words) and compare the correlation with other existing norms of abstractness. For this comparison we use the common subset of two manually and one automatically created resource: MRC Psycholinguistic Database, ratings from Brysbaert et al. (2014) and the automatically created ratings from Turney et al. (2011). We map all existing ratings, as well as our newly created ratings, to the same interval using the method from Köper and Schulte im Walde (2016a). The mapping is performed using a continuous function, that maps the ratings to an interval ranging from very abstract (0) to very concrete (10). The common subset contains 3 665 ratings. Figure 1 shows the resulting pairwise correlation between all four resources. Despite being created automatically, we see that the newly created ratings provide a high correlation with both manually created collections (ρ for MRS=.91, Brysbaert=.93). In addition, the vocabulary of our ratings is much larger than any existing database. Thus this new collection might

[1] `http://nlp.stanford.edu/projects/glove/`
`https://code.google.com/archive/p/word2vec/`

[2] `NN` Implementation based on `https://github.com/amten/NeuralNetwork`

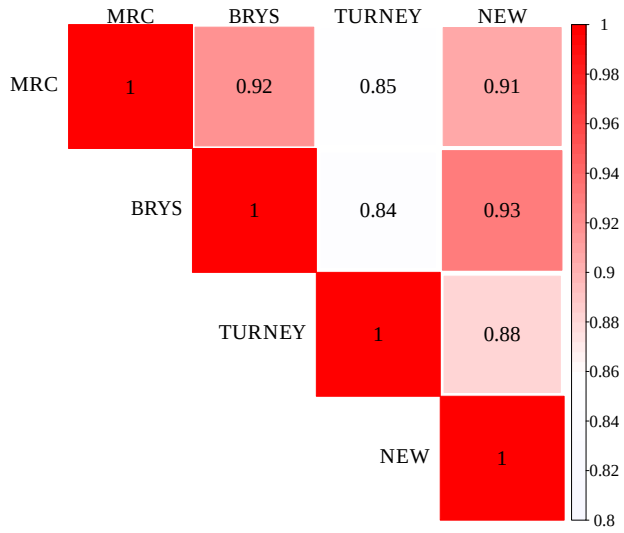

Figure 1: Pairwise Spearman's ρ on commonly covered subset. Red = high correlation.

be useful, especially for further research which requires large vocabulary coverage.[3]

2.2 Abstractness for Phrases

A potential advantage of our method is that abstractness can be learned for multi-word units as long as the representation of these units live in the same distributional vector space as the words required for the supervised training.

In this section we explore if ratings propagated to verb-noun phrases provide useful information for metaphor detection. As dataset we relied on the collection from Saif M. Mohammad and Turney (2016), who annotated different senses of WordNet verbs for metaphoricity (Fellbaum, 1998).

We used the same subset of verb–direct object and verb–subject relations as used in Shutova et al. (2016). As preprocessing step we concatenated verb-noun phrases by relying on dependency information based on a web corpus, the ENCOW14 corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015). We removed words and phrases that appeared less than 50 times in our corpus, thus our selection covers 535 pairs, 238 of which were metaphorical and 297 literal.

Given a verb-noun phrase, such as stamp_person, we obtained vector representations using *word2vec* and the same hyper-parameters that were used for the *W2V300* embeddings (Section 2.1.1) together with the best learning method (NN). The technique allows us to propagate abstractness to every vector, thus we learn abstractness ratings for all three constituents: verb, noun and the entire phrase.

For the metaphor classification experiment we use the rating score and apply the Area Under Curve (AUC) metric. AUC is a metric for binary classification. We assume that literal instances gain higher scores (= more concrete) than metaphorical word pairs. AUC considers all possible thresholds to divide the data into literal and metaphorical. In addition to the rating score we also show results based on cosine similarity and feature combinations (Table 2).

Feat.	Name	Type	AUC
-	Random	baseline	.50
1	V-NN	cosine	.75
2	V-Phrase	cosine	.70
3	NN-Phrase	cosine	.68
4	V	rating	.53
5	NN	rating	**.78**
6	Phrase	rating	.71
Comb	1+2+3	cosine	.75
Comb	4+5+6	rating	.74
Comb	all(1-6)	mixed	.80
Comb	1+5+6	best	**.84**

Table 2: AUC Score single features and combinations. Classifying literal and metaphorical phrases based on the Saif M. Mohammad and Turney (2016) dataset.

As shown in Table 2, the rating of the verb alone (AUC=.53) provides almost no useful information. The best performance based on a single feature is the abstractness value of the noun (.78) followed by the cosine between verb and noun vector representation (.75). The phrase rating alone performs moderate (.71). However when combining features we found that the best combinations are obtained by integrating the phrase rating. In more detail, combining noun and phrase rating (5+6) obtains a AUC of (.80). When adding the cosine (1) we obtain the best score of (.84). For comparison, the verb plus noun ratings (4+5) obtains a lower score (.72), this shows that the phrase rating provides complementary and useful information.

[3] Ratings available at http://www.ims.uni-stuttgart.de/data/en_abst_norms.html

2.3 Sense-specific Abstractness Ratings

In this section we investigate if automatically learned multi-sense abstractness ratings, that is having different ratings per word sense, are potentially useful for the task of metaphor detection.

Recent advances in word representation learning led to the development of algorithms for non-parametric and unsupervised multi-sense representation learning (Neelakantan et al., 2014; Liu et al., 2015; Li and Jurafsky, 2015; Bartunov et al., 2016). Using these techniques one can learn a different vector representation per word sense. Such representations can be combined with our abstractness learning method from section 2.1.1.

While in theory any multi-sense learning technique can be applied, we decided for the one introduced by Pelevina et al. (2016), as it performs sense learning after single senses have been learned. Starting from the public W2V300 representations we apply the multi-sense learning technique using the default settings and learn sense-specific word representations. Finally we propagate abstractness to every newly created sense representation by using the exact same model and training data as in Section 1. For a given word in a sentence we can now disambiguate the word sense by comparing its sense-specific vector representation to all context words. The context words are represented using the (single sense) global representation. We always pick the sense representation that obtains the largest similarity, measured by cosine. The potential advantage of this method is that in a metaphor detection system we are now able to look up word-sense-specific abstractness ratings instead of globally obtained ratings.

For this experiment we use the VU Amsterdam Metaphor Corpus (Steen, 2010) (VUA), focusing on verb metaphors. The collection contains 23 113 verb tokens in running text, annotated as being used literally or metaphorically. In addition we present results for the TroFi metaphor dataset (Birke and Sarkar, 2006) containing 50 verbs and 3 737 labeled sentences. We pre-processed both recourses using *Stanford CoreNLP* (Manning et al., 2014) for lemmatization, part-of-speech tagging and dependency parsing.

We present results by applying ten-fold cross-validation over the entire data. For the VUA we additionally present results for the test data using the same training/test split as in Beigman Klebanov et al. (2016).

Abstractness norms are implemented using the same five feature dimensions as used by Turney et al. (2011) plus dimensions respectively for subject and object, thus we rely on the seven feature, namely:

1. Rating of the verbs subject
2. Rating of the verbs object
3. Average rating of all nouns (excluding proper names)
4. Average rating of all proper names
5. Average rating of all verbs, excluding the target verb
6. Average rating of all adjectives
7. Average rating of all adverbs

For classification we used a balanced Logistic Regression classifier following the findings from Beigman Klebanov et al. (2015). While this default setup tries to generalize over unseen verbs by only looking at a verb's context we further present results for a second setup that uses a 6th feature: namely the lemma of the target verb itself (+L). The purpose of the second system is to describe performance with respect to the state of the art (Beigman Klebanov et al., 2016), which among other features also uses the verb lemma.

Feat.	TroFi(10F)	VUA(10F)	VUA(Test)
1S	.72	.42	.44
MS	**.74**	**.44***	**.46**
1S(+L)	.74	**.61**	**.62**
MS(+L)	**.75**	**.61**	**.62**

Table 3: F-score (Metaphor). Classifying literal and metaphorical verbs based on the VUA and TroFi dataset. MS = multi-sense, 1S= single sense.

As shown in Table 3, the mutli-sense ratings constantly outperform the single-sense ratings in a direct comparison on all three sets. The difference in performance of single and multi-sense ratings is statistically significant on the full VUA dataset, using the χ^2 test and $*$ for $p < 0.05$. However we also notice that the effect vanishes as soon as we combine the ratings with the lemma of the verb, which is especially the case for the VUA dataset where the lemma increases the performance by a large margin. In contrast to related work, the system with the verb unigram (+UL)

can be considered state-of-the-art. When applying the same evaluation as Beigman Klebanov et al. (2016), namely a macro-average over the four genres of VUA, we obtain an average f-score of .60 by using only eight feature dimensions and abstractness ratings as external resource.[4]

3 Conclusion

In this paper we compared supervised methods to propagate abstractness norms to words. We showed that a neural-network outperforms other methods. In addition we showed that norms for multi-words phrases can be beneficial for type based metaphor detection. Finally we showed how norms can be learned for sense representations and that sense specific norms show a clear tendency to improve token-based verb metaphor detection.

Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732 (Maximilian Köper) and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde).

References

Lawrence W Barsalou and Katja Wiemer-Hastings. 2005. Situating Abstract Concepts. *Grounding cognition: The role of perception and action in memory, language, and thought*, pages 129–163.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking Sticks and Ambiguities with Adaptive Skip-gram. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 130–138, Cadiz, Spain.

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised Word-level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20, Denver, Colorado.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic Classifications for Detection of Verb Metaphors. In *Proceedings of the h Annual Meeting of the Association for Computational Linguistics*, pages 101–106, Berlin, Germany.

Julia Birke and Anoop Sarkar. 2006. A Clustering Approach for the Nearly Unsupervised Recognition of Nonliteral Language. In *Proceedings of the 11th Conference of the European Chapter of the ACL*, pages 329–336, Trento, Italy.

[4]SOA results from Klebanov obtain also a score of .60

Marc Brysbaert, AmyBeth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand Generally known English Word Lemmas. *Behavior Research Methods*, pages 904–911.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-Level Metaphor Detection using Neural Networks. In *Proceedings of The Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, CA, USA.

Jonathan Dunn. 2013. What Metaphor Identification Systems can tell us About Metaphor-in-language. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 1–10, Atlanta, Georgia.

Jonathan Dunn. 2015. Modeling Abstractness and Metaphoricity. *Metaphor and Symbol*, pages 259–289.

Christiane Fellbaum. 1998. A Semantic Network of English Verbs. In Christiane Fellbaum, editor, *WordNet – An Electronic Lexical Database*, Language, Speech, and Communication. MIT Press, Cambridge, MA.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. 2015. Distributional Vectors Encode Referential Attributes. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal.

Hessel Haagsma and Johannes Bjerva. 2016. Detecting Novel Metaphor using Selectional Preference Information. In *Proceedings of The Fourth Workshop on Metaphor in NLP*, pages 10–17, San Diego, CA, USA.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, Dave Barner, Donald Black, Majorie Friedman, and Ralph Weischedel. 2013. Automatic Extraction of Linguistic Metaphors with LDA Topic Modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66, Atlanta, Georgia.

Felix Hill, Anna Korhonen, and Christian Bentz. 2014. A Quantitative Empirical Analysis of the Abstract/Concrete Distinction. *Cognitive Science*, 38:162–177.

Richard Kammann and Lynn Streeter. 1971. Two Meanings of Word Abstractness. *Journal of Verbal Learning and Verbal Behavior*, 10(3):303 – 306.

Maximilian Köper and Sabine Schulte im Walde. 2016a. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2595–2598, Portoroz, Slovenia.

Maximilian Köper and Sabine Schulte im Walde. 2016b. Distinguishing Literal and Non-Literal Usage of German Particle Verbs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, California.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago Press.

Jiwei Li and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1722–1732, Lisbon, Portugal.

Linlin Li and Caroline Sporleder. 2009. Classifier Combination for Contextual Idiom Detection Without Labelled Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 315–323.

Linlin Li and Caroline Sporleder. 2010. Using Gaussian Mixture Models to Detect Figurative Language in Context. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2015. Learning Context-Sensitive Word Embeddings with Neural Tensor Skip-Gram Model. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1284–1290, Buenos Aires, Argentina.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

James H. Martin. 1996. Computational Approaches to Figurative Language. *Metaphor and Symbolic Activity*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1059–1069, Doha, Qatar.

Maria Pelevina, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making Sense of Word Embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183, Berlin, Germany, August.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar.

Ekaterina Shutova Saif M. Mohammad and Peter D. Turney. 2016. Metaphor as a Medium for Emotion: An Empirical Study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*Sem)*, pages 23–33, Berlin, Germany.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, and Andreas Witt, editors, *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–24, Lancaster.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor Corpus Annotated for Source - Target Domain Mappings. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 3225–3261, Valletta, Malta.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor Identification Using Verb and Noun Clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170.

G. Steen. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging Evidence in Language and Communication Research. John Benjamins Publishing Company.

James H Steiger. 1980. Tests for Comparing Elements of a Correlation Matrix. *Psychological Bulletin*.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual Metaphor Detection Using Common Semantic Features. In *Proceedings of the Workshop on Metaphor in NLP*, pages 45–51, Atlanta, USA.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258.

Peter D. Turney and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, pages 315–346.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

Gabriella Vigliocco, Stavroula-Thaleia Kousta, Pasquale Anthony Della Rosa, David P Vinson, Marco Tettamanti, Joseph T Devlin, and Stefano F Cappa. 2014. The Neural Representation of Abstract Words: the Role of Emotion. *Cerebral Cortex*, pages 1767–1777.