Verbal Idioms: Concrete Nouns in Abstract Contexts

Jean Charbonnier
Hochschule Hannover
Expo Plaza 12
30539 Hannover
Jean.Charbonnier@hs-hannover.de

Christian Wartena
Hochschule Hannover
Expo Plaza 12
30539 Hannover
Christian.Wartena@hs-hannover.de

Abstract

In this paper, we present our approach for the *KONVENS 2021* shared task *Disambiguation of German Verbal Idioms*. Our model is a decision tree-based classifier that uses static word embeddings and computed concreteness values to predict whether a verbal idiom is used figuratively or literal.

1 Introduction

The task of the competition *Disambiguation of German Verbal idioms* was to determine whether a verbal idiom in a given context is used in its idiomatic figurative or metaphorical meaning or whether the literal meaning is intended. In a few cases both meanings are intended and in some other cases the correct meaning cannot be chosen from the context. We will nevertheless consider the task as a binary classification problem and discard the underrepresented classes for our evaluation.

The main idea that we pursue is that an expression that is used metaphorically is not used it its typical context. Consider example (1). Here the idiomatic expression *die Notbremse ziehen* (pull the emergency break) is used literally. An emergency break is usually found in trains, and this example contains a lot of words from the rail domain and words denoting train furniture.

(1) Fenster, Türen und Sitze wurden als Ausdruck der Vorfreude beschädigt. Weil offenbar übermüdete Raver in einem Sonderzug aus Hannover fünfmal die Notbremse gezogen hatten, schnitten sie sich ins eigene Fleisch und kamen mit einer Stunde Verspätung in Berlin an. Und dann gab es noch Flaschenwürfe von Ravern auf einem Fahrgastschiff an der Mühlendammbrücke.

(Windows, doors and seats were damaged as an expression of anticipation. Because apparently overtired ravers had pulled the emergency brake five times on a special train from Hanover, they cut themselves into their own flesh and arrived in Berlin an hour late. And then there were still bottles thrown by ravers on a passenger ship at the Mühlendamm Bridge.)

In example (2) the same expression is used figuratively and no other words from the rail domain are found.

(2) Der gemeinsame Entwurf für eine modifizierte Fristenregelung von SPD- und FDP-Frauen war erstaunlich weit gediehen. Zu weit für die Männer in der FDP, die in letzter Sekunde die Notbremse zogen. Der Koalitionspartner CDU könne verärgert reagieren, hieß im Vorstand.

(The joint draft for a modified period regulation by SPD and FDP women had come astonishingly far. Too far for the men in the FDP who pulled the emergency brake at the last second. The coalition partner CDU could react angrily, said the board.)

With carefully selected examples this sounds plausible, but in real life there are a number of challenges: (1) the figurative meaning can be used in its literal context as well, and many authors like to do so; (2) sometimes the metaphor can be extended with more terms from the domain of the literal meaning. (3) It is not trivial to find words that are typical for the original (literal) domain of the expression. Since the expressions in the task are idiomatic, they are more often used in their figurative than in their literal meaning. If we collect typical contexts (e.g.

in the form of word embeddings) of these words, we include many words typical for the metaphoric use of the word.

To overcome this last problem, we use concreteness as a proxy to distinguish typical contexts of the word in its literal meaning from other contexts. In almost all cases, the noun in the idiomatic verbal expression refers to a very concrete thing, and thus is called a concrete word (Spreen and Schulz, 1966; Friendly et al., 1982; Brysbaert et al., 2014). As was noted by several studies (see e.g. (Tanaka et al., 2013)) and investigated in detail by Frassinelli and Schulte im Walde (2019), concrete words tend to occur in the context of other concrete words. Thus, if we find many abstract words and only a few concrete words in the context of the idiomatic expression, it is very likely that those words are not from the original domain of the expression and the expression thus is used figuratively.

Thus, we use the concreteness of the context of the expression and the semantic similarity of the expression with its context as features to train a classifier.

The remainder of this paper is organized as follows: First, we discuss related work. This is followed by a description of the data set provided for this challenge and an description of our method, here we also lay out an analysis of our used features. In Section 5 we show our results on the test set and during our cross-validation. Finally, in Section 6 we share our insights.

2 Related Work

Concreteness of words has been studied for several decades in the field of psycho-linguistics 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. In (Charbonnier and Wartena, 2020, 2019) we show a method to obtain concreteness predictions using static word embeddings. We trained a regression model to predict the concreteness of words based on their word embedding using data gather by humans for training. This method was also used for this paper.

Concreteness has been used for detection of metaphors and non-literal language before by e.g. Turney et al. (2011), who interpreted the task as a disambiguation between the literal and metaphorical sense of a word. They used logistic regression for this binary classification task in which the concreteness values of both the adjective and the noun are the main features. Hill and Korhonen (2014) also analyzed adjective-noun pairs, using the average concreteness of these pairs to decide whether the literal sense of the pair is metaphorical or not. Frassinelli and Schulte im Walde (2019) analyzed the cooccurrences of abstract and concrete nouns and verbs. They found consistent patterns in the distributional representation of subcategorising and subcategorised concrete and abstract word.

Katz and Giesbrecht (2006) used dense word vectors to discriminate between literal and idiomatic occurrences of the German single verbal idiom *ins Wasser fallen*[1]. Latent Semantic Analysis and singular value decomposition were used to construct a vector representations for the compositional and non-compositional interpretation of this multi word expression. They reported that the two vectors where nearly orthogonal in this example and concluded that the contexts must be highly different for the two versions.

Cook et al. (2007) found that the syntactic configurations of an idiom often gives its meaning away. They used the canonical form of the verbal idiom, which are its most preferred syntactic patterns as a feature.

Ehren et al. (2020) used a 2-layer LSTM network to get latent representations for the verbal idiom tokens. These were then used in a fully connected layer to predict the class using softmax. They used pretrained static[2] and contextualized[3] word embeddings as an input for their model.

Li and Sporleder (2009) and Ehren (2017) proposed methods inspired by cohesion-based graphs to discriminate between literal and non-literal use of non-compositional multiword expressions. It is based on the assumption that words in a text form a cohesive unit and if this cohesion is weakened by an expression it is classified as literal, otherwise its assumed to be metaphorical. The cohesiveness of a text is measured using said cohesion graph using word embedding and cosine similarity.

3 Data

The dataset is provided by the task organizers (Ehren et al., 2021). The complete data set is a merge of *COLF-VID* (Ehren et al., 2020) and the *German SemEval-2013 task 5b* (Korkontzelos et al.,

[1] Literal: To fall into water. Idiomatic: to be canceled
[2] fastText (Bojanowski et al., 2016)
[3] ELMo (Peters et al., 2018)

2013). The training data consist of 6902 short text fragments (usual 3 sentences) containing a verbal idiom in the middle sentence that is marked in the text using <b>and <\b> tags for each word of the idiom. (Ehren et al., 2021) The data is provided as a TSV-file with the following format:

`ID \t Verbal-Idiom \t Label \t Text`

The test data contains of 1511 examples in exactly the same format as the training data. 264 of these examples contain verbal idioms that were not part of the training data, together 3 new unique verbal idioms. These are *mit Feuer spielen* (84), *Korb bekommen* (94) and *Frucht tragen* (90).

4 Method

To predict whether an idiom is either being classified as figuratively or metaphorically, we trained a Random Forest Classifier (Ho, 1995) using the features described in Section 4.1.

4.1 Features

We assume that a verbal idiom is likely to be literal if many related words are found in the context and if also many concrete words are found in the same context. Moreover, we assume that the properties of a close context are more important than those of the wider context. Thus we divide the context in different parts to get different views of context. To find the closest context of the verbal idiom, we use the dependency parser of SpaCy (Honnibal et al., 2020). We now define the following regions (see Figure 1 for an example):

subject The subject of the verb in the verbal idiom. In case that the verb is used as an attributive present participle, we take the modified noun as the subject.

object The object of the verb in the verbal idiom (if present).

phrase The smallest constituent that contains all parts of the verbal idiom.

sentence The sentence containing the verbal idiom.

text The whole text provided (usually three sentences).

In all cases, we remove the words belonging to the verbal idiom and all stopwords.

To compute the similarity measures, we use pretrained fastText embeddings (Bojanowski et al., 2016) and compute each time the cosine between the average vector of all words in the verbal idiom and all nouns in the considered part of the context region as described before.

In fact, this similarity is not exactly what we would need, since we only want to see whether words are from the same domain like railways, fishery, furniture, etc. Since we have no clear definition of what these domains are exactly and what domains are relevant, we construct a term document matrix and use LSA (Landauer et al., 1998) to find 20 latent topics. Now we extract for each word a vector with a weight for each of the 20 topics. We use these vectors again to compute the similarity between the verbal idiom and each piece of context. This time all words (not just the nouns) are included in the average vector. If the subject or object is missing or contains no nouns the average of all word embeddings is used for that phrase instead.

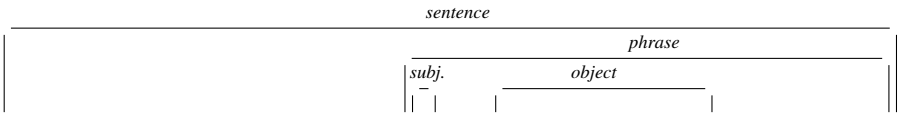
For the concreteness of each part of the context we use again only the nouns and consider both the average concreteness score and the maximum concreteness score[4]. The idea behind using the maximum is that in a complex noun phrase, the most concrete noun determines the concreteness of the whole phrase. By modifying a noun, it usually becomes more specific and more concrete. In case a part of the context is not present, not found or empty after stopwords and words from the verbal idiom are removed, we use the average concreteness of all words as default. All combinations of concreteness and similarity measures with different text regions give us in total 20 features. An overview of the features is given in Table 1 that shows the importance of each feature.

4.2 Classifiers and one Additional Feature

We use the RandomForest implementation from scikit-learn (Pedregosa et al., 2011) as our classifier to learn a model from the 20 features described above. We also introduced class weights: 0.8 and 0.2 for literally and figuratively, respectively. From our tests, this seems to be a good value to give the under represented class of literally idioms a boost, while not overfitting the model. For the same reason we set the maximum depth of the tress to 9.

Since the optimal value for a feature to split be-

[4]We describe the method to to gather the concreteness scores in detail in (Charbonnier and Wartena, 2019)



. . . führte. Im Jahre 1978 wähnte die Berlin Consult, sie habe ein lukratives Geschäft ***an Land gezogen****. Für Saudi- . . .*

Figure 1: Example of the different regions that are considered to construct similarity and concreteness features. The region called *text* also includes the sentences before and after the sentence with the idiom. In all cases, the bold words, that are part if the verbal idiom, are excluded from the similarity and concreteness computation.

tween the two classes might be different for each verbal idiom, we also add an identifier for the verbal idiom as a feature. We consider two variants: one using this feature and one without knowledge about the verbal idiom that is classified.

4.3 Feature Analysis

In a number of cases there is no subject or no object, or the subject, object or phrase could not be identified, or it was found but did not contain any nouns that are not part of the verbal idiom. In these cases, a default value was used. For the subject in 27% of the cases a default value was used, for objects in 58% of the cases and for phrases in 2% . In a few cases, we also do not have a word embedding for proper nouns. In this case, we also use the default embedding.

In order to get some insight in the strength and usefulness of the single features, we used the Mean Decrease of Impurity (MDI) (Breiman et al., 1984) to get the importance of the single features. The MDI calculates each features importance as the sum over the number of splits that include a feature, weighted by the number of samples it splits, averaged over all trees of the ensemble. In Table 1 we show the single features combinations with their MDI values, in Figure 2 we show all features together with their standard deviation over the different estimators inside the Random Forest. When we look at the parts of the text, we see that the whole text is the most important region, especially for the concreteness measures. The subject and the object based features have the smallest MDI values, which might be explained by the high number of default values used here. The most outstanding observations from the MDI analysis are the high values for the concreteness features. Finally, average concreteness seems to better separate the classes than the maximum concreteness.

We also have a look (Figure 3) at the correlation between pairs of features, since we expect that many features almost have the same information. Indeed, we see a very high correlation as expected between some pairs of features, like between the average and maximum concreteness of the phrase. In many cases only one noun is found in the phrase (besides the one belonging to the idiom) and the average and maximum concreteness is the same. A number of other high correlation values can be explained again by the high fraction of default values. Remarkably, we also find a number of negative correlation. In almost all cases this involves a pair of features from which one is based on the subject and one on the object. Given the high fraction of default values for subject and objects we are reticent to draw any conclusions from this.

4.4 Experiments

During the training phase we used 10-fold cross-validation to test the robustness of our method further. We used two ways to divide the the training data in 10 parts: in the stratified sampling we divided the data in 10 parts such that each part contained the same number of literal and figurative sentences for each idiom. Thus, in this condition there is most likely never an unseen idiom in the test data. In the ordered sampling, we just take the first 690 examples, the following 690 examples and so on. Thus, each time the test data contain a large number of unseen idioms and a number of idioms for which only very few instances were seen in the training data.

In the following section we report on results from both conditions as well as on the results on the official test data.

5 Results

Table 2 presents the results of our experiments. We evaluated two groups: *all verbal idioms* and only *unseen verbal idioms*. The later is composed of all instances where verbal idioms occur in the test set during cross-validation but not in the training set. Our results are in line with the unseen verbal idioms we encounter in the shared tasks test

Table 1: Mean decrease of impurity (MDI) of each feature indicating the importance of each feature to detect literal use of a verbal idiom. In addition for each row and each column the sum of the MDI-values is given to get an impression of the importance of each type and group of features.

	subj.	obj.	phrase	sent.	text	Σ
LSA sim.	0.020	0.009	0.030	0.024	0.038	0.121
FT sim.	0.032	0.030	0.060	0.032	0.041	0.195
max concr.	0.019	0.014	0.069	0.066	0.069	0.237
avg concr.	0.027	0.015	0.112	0.118	0.173	0.445
Σ	0.098	0.068	0.271	0.240	0.321	

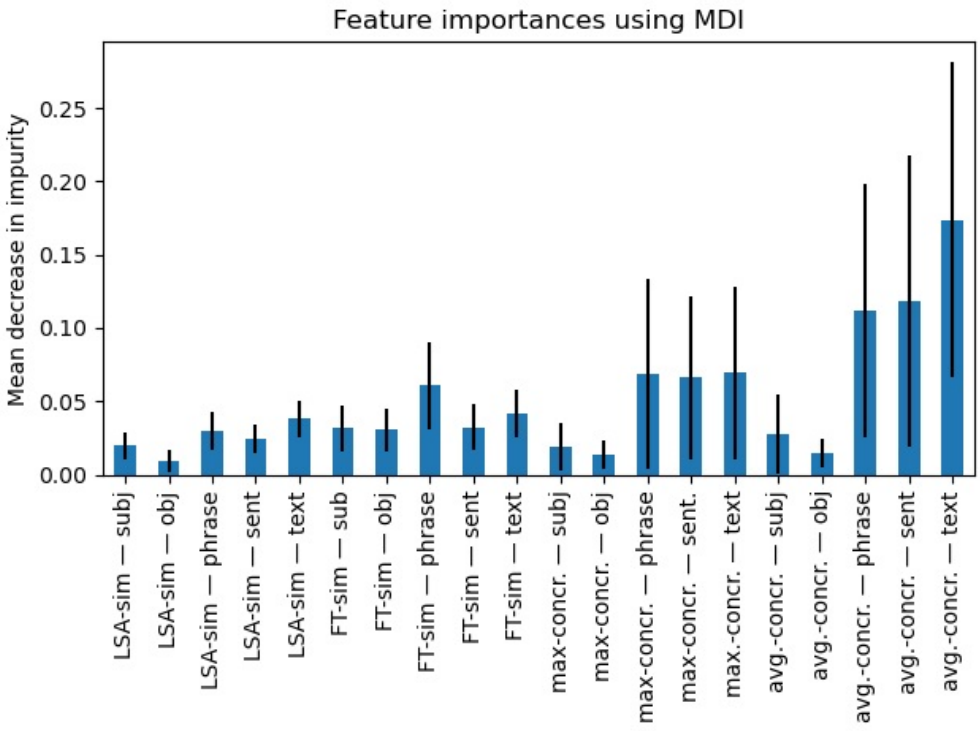


Figure 2: Mean decrease of impurity (MDI) for each feature, showing the usefulness of each feature to be used in a decision tree.

set. Adding the information about the identity of the verbal idiom does not help the classifier much. It performs slightly better on known idioms, also show good results using cross validation, but interestingly much worse on unknown idioms in the test set. This is most likely due to the fact that decision trees do not work well with categorical data as they cannot find a clear boundary and the fact that the test set is not an averaged result like the result for our cross-validation splits.

6 Discussion

As expected the results for the stratified cross-validation are slightly better than for the ordered cross-validation. Especially, the recall turns out to be significantly better in the stratified experiments. Clearly, the classifier learns something about specific idioms and does not generalize enough. Remarkably, if the verbal idiom itself is used as a feature, the results for the unseen idioms also improve in the ordered cross validation. In three cases, the results for the official test set fall in between those of the ordered and stratified cross-validation. In one case there is a large difference between the results: the classification of the unseen verbal idioms in the official test performed significantly worse than the classification of the unseen idioms using cross-validation in case the idiom is added as a feature.

Our conjecture was that we can detect figurative use of verbal idioms by the degree to which they fit in the context. Whether the verbal idiom consists of

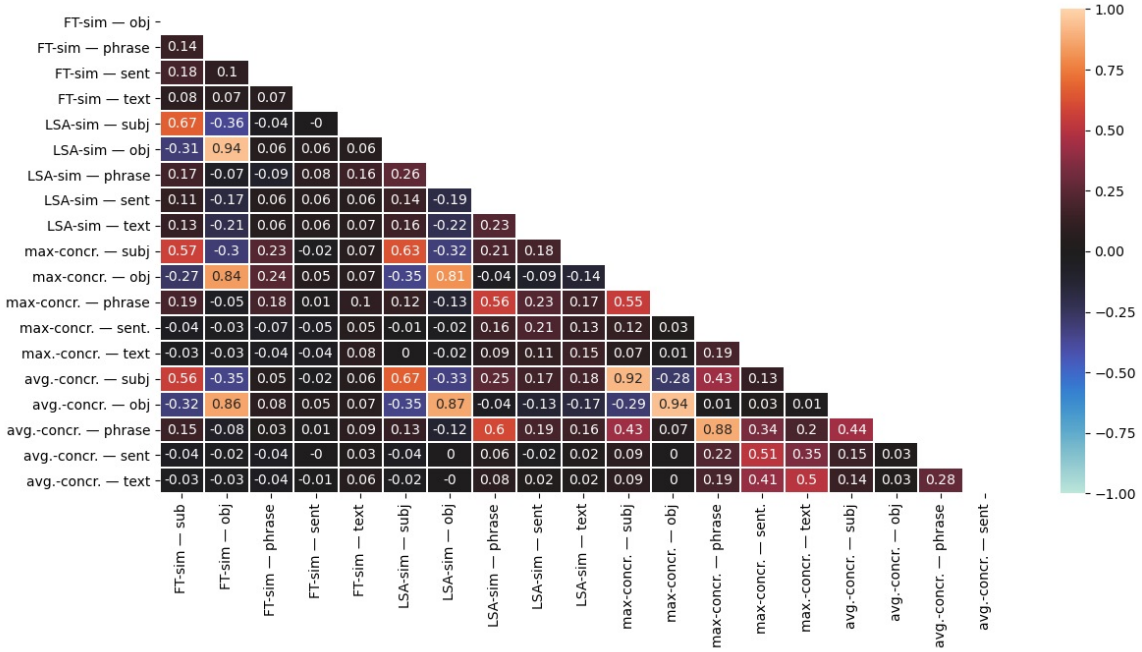


Figure 3: Pearson's correlation between each pair of features.

Table 2: Results

	all verbal idioms			unseen verbal idioms		
	prec.	recall	F1	prec.	recall	F1
20 features						
cross val. (ordered)	0.528	0.538	0.511	0.408	0.501	0.417
cross val. (stratified)	0.596	0.613	0.603	-	-	-
test set	0.600	0.576	0.588	0.425	0.415	0.420
20 features + verbal idiom						
cross val. (ordered)	0.571	0.507	0.523	0.494	0.502	0.452
cross val. (stratified)	0.602	0.627	0.614	-	-	-
test set	0.596	0.617	0.607	0.857	0.146	0.250

words from the same domain or words related otherwise to the words in the rest of the sentence could be measured by the similarity of their distributional representations. Since verbal idioms are usually used in their figurative meaning, the picture arising from this comparison might not be as clear as we would like. Since usually very concrete words, i.e. words denoting things that easily can be perceived by the senses, are used in metaphors to talk about abstract things, we expect that metaphoric use of a verbal idiom is mainly found in sentences with a lot of abstract concepts and literal use in contexts with many concrete words. This conjecture is indeed affirmed by the experiment: we get acceptable results and the features based on concreteness are the most important ones. This finding is also in line with other experiments on other cases of metaphoric language reported in the literature.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Leo Breiman, Jerome Friedman, Richard A. Olshen, and Charles. J. Stone. 1984. *Classification and Regression Trees*. Taylor & Francis.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on*

Computational Semantics - Long Papers, pages 176–187, Gothenburg, Sweden. Association for Computational Linguistics.

Jean Charbonnier and Christian Wartena. 2020. Predicting the concreteness of german words. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, volume 2624 of *CEUR Workshop Proceedings*.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 41–48, Prague, Czech Republic. Association for Computational Linguistics.

Rafael Ehren. 2017. Literal or idiomatic? identifying the reading of single occurrences of German multiword expressions using word embeddings. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–112, Valencia, Spain. Association for Computational Linguistics.

Rafael Ehren, Laura Kallmeyer, Tim Lichte, and Jakub Waszczuk. 2021. Shared task on the disambiguation of german verbal idioms. `https://github.com/rafehr/vid-disambiguation-sharedtask`.

Rafael Ehren, Timm Lichte, Laura Kallmeyer, and Jakub Waszczuk. 2020. Supervised disambiguation of german verbal idioms with a bilstm architecture. In *Fig-Lang@ACL*.

Diego Frassinelli and Sabine Schulte im Walde. 2019. Distributional interaction of concreteness and abstractness in verb–noun subcategorisation. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 38–43, Gothenburg, Sweden. Association for Computational Linguistics.

Michael Friendly, Patricia E. Franklin, David Hoffman, and David C. Rubin. 1982. 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.

Felix Hill and Anna Korhonen. 2014. Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–731.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multiword expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Linlin Li and Caroline Sporleder. 2009. A cohesion graph based approach for unsupervised recognition of literal and non-literal use of multiword expressions. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, pages 75–83, Suntec, Singapore. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Otfried Spreen and Rudolph W. Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning & Verbal Behavior*, 5(5):459–468.

Shinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi 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, pages 475–484, New York, NY, USA. ACM.

Peter D. 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*, EMNLP '11, pages 680–690, Stroudsburg, PA, USA. Association for Computational Linguistics.