# Do Character-Level Neural Network Language Models Capture Knowledge of Multiword Expression Compositionality?

**Ali Hakimi Parizi** and **Paul Cook**
Faculty of Computer Science, University of New Brunswick
Fredericton, NB E3B 5A3 Canada
ahakimi@unb.ca, paul.cook@unb.ca

## Abstract

In this paper we propose the first model for multiword expression (MWE) compositionality prediction based on character-level neural network language models. Experimental results on two kinds of MWEs (noun compounds and verb-particle constructions) and two languages (English and German) suggest that character-level neural network language models capture knowledge of multiword expression compositionality, in particular for English noun compounds and the particle component of English verb-particle constructions. In contrast to many other approaches to MWE compositionality prediction, this character-level approach does not require token-level identification of MWEs in a training corpus, and can potentially predict the compositionality of out-of-vocabulary MWEs.

## 1 Introduction

Multiword expressions (MWEs) are lexical items that are composed of multiple words, and exhibit some degree of idiomaticity (Baldwin and Kim, 2010), for example semantic idiomaticity, in which the meaning of an MWE is not entirely transparent from the meanings of its component words, as in *spill the beans*, which has an idiomatic meaning of 'reveal a secret'. Compositionality is the degree to which the meaning of an MWE is predictable from the meanings of its component words. It is typically viewed as lying on a continuum, with expressions such as *speed limit* and *gravy train* lying towards the compositional and non-compositional ends of the spectrum, respectively, and expressions such as *rush hour* and *fine line* falling somewhere in between as semi-compositional.[1] Compositionality can also be viewed with respect to an individual component word of an MWE, where an MWE component word is compositional if its meaning is reflected in the meaning of the expression. For example, in *spelling bee* and *grandfather clock*, the first and second component words, respectively, are compositional, while the others are not.

Knowledge of multiword expressions is important for natural language processing (NLP) tasks such as parsing (Korkontzelos and Manandhar, 2010) and machine translation (Carpuat and Diab, 2010). In the case of translation, compositionality is particularly important because a word-for-word translation would typically be incorrect for a non-compositional expression. Much research has therefore focused on compositionality prediction of MWEs, primarily at the type level. One common approach to measuring compositionality is to compare distributional representations of an MWE and its component words (e.g., Schone and Jurafsky, 2001; Baldwin et al., 2003; Katz and Giesbrecht, 2006; Reddy et al., 2011; Schulte im Walde et al., 2013; Salehi et al., 2015). The hypothesis behind this line of work is that the representation of a compositional MWE will be more similar to the representations of its component words than the representation of a non-compositional MWE will be to those of its component words. One issue faced by such approaches is that token-level instances of MWEs must be identified in a corpus in order to form distributional representations of them. Token-level MWE identification has been studied for specific types of MWEs such as verb-particle constructions (e.g., Kim and Baldwin, 2010) and



[1] These expressions and compositionality judgements are taken from Reddy et al. (2011).



*Proceedings of the Joint Workshop on*
*Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 185–192
Santa Fe, New Mexico, USA, August 25-26, 2018.

verb–noun idioms (e.g., Salton et al., 2016). Broad coverage MWE identification has also been studied, and remains a challenge (Schneider et al., 2014; Gharbieh et al., 2017).

Language models are common throughout NLP in tasks including machine translation (Brants et al., 2007), speech recognition (Collins et al., 2005), and question answering (Chen et al., 2006). Although word-level language models are widely used, and their performance can be higher than character-level language models, character-level models have the advantage that they can model out-of-vocabulary words (Mikolov et al., 2012). Owing to this advantage, character-level language models have been applied in a range of NLP tasks, including authorship attribution, (Peng et al., 2003), part-of-speech tagging (Santos and Zadrozny, 2014), case restoration (Susanto et al., 2016), and stock price prediction (dos Santos Pinheiro and Dras, 2017). Moreover, character-level information can be composed to form representations of words (Ling et al., 2015).

In this paper we consider whether character-level neural network language models capture knowledge of MWE compositionality. We train character-level language models based on recurrent neural networks — including long short-term memory (LSTM, Hochreiter and Schmidhuber, 1997) and gated recurrent unit (GRU, Cho et al., 2014). We then use these language models to form continuous vector representations of MWEs and their component words. Following prior work, we then use these representations to predict the compositionality of MWEs. This method overcomes the limitation of previous work in this vein of having to identify token instances of MWEs in a corpus in order to form a distributional representation of them. Moreover, this approach could potentially be applied to predict the compositionality of out-of-vocabulary expressions that were not seen in the corpus on which the language model was trained. To the best of our knowledge, this is the first work to apply character-level neural network language models to predict MWE compositionality. Our experiments on two kinds of MWEs (noun compounds and verb-particle constructions) and two languages (English and German) produce mixed results, but suggest that character-level neural network language models do indeed capture some knowledge of multiword expression compositionality, in particular for English noun compounds and the particle component of English verb-particle constructions.

## 2 A Character-level Model for MWE Compositionality

If an MWE is compositional, it is expected to be similar in meaning to its component words. Since the vector representation of a word/MWE is taken as a proxy for its meaning, we expect the vector representation of a compositional MWE to be similar to its component words' vectors. In order to obtain vectors representing each of an MWE and its component words through a character-level neural network language model, each of the MWE and its component words are considered as a sequence of characters. Each of these character sequences includes a special end-of-sequence character. In the case of an MWE, the character sequence includes a space character between the component words. For example, the MWE *ivory tower* is represented as the sequence $< i, v, o, r, y, \ , t, o, w, e, r, \text{END} >$. These character sequences are fed to the neural network language model, and the hidden state of the neural network at the end of the sequence is taken as the vector representation for that sequence.[2]

Once vector representations of an MWE and its component words are obtained, following Salehi et al. (2015), the following equations are then used to compute the compositionality of an MWE:

$$\text{comp}_1(\text{MWE}) = \alpha sim(\text{MWE}, \text{C}_1) + (1 - \alpha) sim(\text{MWE}, \text{C}_2) \quad (1)$$

$$\text{comp}_2(\text{MWE}) = sim(\text{MWE}, \text{C}_1 + \text{C}_2) \quad (2)$$

where $\text{MWE}$ is the vector representation of the MWE, and $\text{C}_1$ and $\text{C}_2$ are vector representations for the

[2]This approach does have the limitation that it is not immediately clear how to input a "gappy" MWE, such as *give X a chance*, to the language model. A possible solution would be to attempt to select a prototypical slot filler. However, this issue does not arise in this study because the evaluation datasets used — English and German noun compounds, and English verb-particle constructions — do not consist of gappy expressions. (Although English verb-particle constructions can appear in the split configuration, we input them to the language model in the joined configuration.)

| Language | Number of Characters | Number of Tokens | Size |
|---|---|---|---|
| English | 102M | 16.5M | 103 MB |
| German | 102M | 14.2M | 102 MB |

Table 1: The size of the English and German training corpora in terms of characters, tokens, and megabytes.

first and second components of the MWE, respectively.[3] In both cases, we use cosine as the similarity measure. $\text{comp}_1$ is based on Reddy et al. (2011). As shown in equation (1), the compositionality of an MWE is computed based on measuring the similarity of the MWE and each of its component words, and then combining these two similarities into an overall compositionality score. $\text{comp}_2$ is based on Mitchell and Lapata (2010) and measures compositionality by considering the similarity between the MWE and the summation of its component words' vectors.

## 3 Materials and Methods

In this section, we describe the language model and corpus it was trained on, as well as the evaluation dataset and methodology.

### 3.1 Language Model

We use a publicly available TensorFlow implementation of a character-level RNN language model.[4] We use the following parameter settings as defaults: a two-layer LSTM with one-hot character embeddings and a hidden layer size of 128 dimensions. The batch size, learning rate, and dropout are set 20, 0.002, and 0, respectively.[5] We consider some alternative parameter settings to these defaults in section §4.

### 3.2 Training Corpus

We train language models over a portion of English and German Wikipedia dumps — following Salehi et al. (2015) — from 20 January 2018. The raw dumps are preprocessed using WP2TXT[6] to remove wikimarkup, metadata, and XML and HTML tags.

The text from Wikipedia contains many characters that are not typically found in MWEs, for example, non-ASCII characters. Such characters drastically increase the size of the vocabulary of the language model, which leads to very long training times. We therefore remove all non-ASCII characters from the English dump, and all non-ASCII characters other than *ä, Ä, ö, Ö, ü, Ü, ß* from the German dump.

Training the character-level language model over the Wikipedia dumps in their entirety would take a prohibitively long time due to their size. We therefore instead carry out experiments training on a 1% sample of the English dump, and a 2% sample of the German dump (to give a corpus of similar size to the English one). Details of the resulting training corpora are provided in table 1.

### 3.3 Evaluation Data

The proposed model is evaluated over the same three datasets as Salehi et al. (2015), which cover two languages (English and German) and two kinds of MWEs (noun compounds and verb-particle constructions).

**ENC** This dataset contains 90 English noun compounds (e.g., *game plan*, *gravy train*) which are annotated on a scale of [0,5] for both their overall compositionality, and the compositionality of each of their component words (Reddy et al., 2011).

[3] Although $\text{comp}_1$ and $\text{comp}_2$ are formulated for MWEs with two component words, they could be extended to handle MWEs with more than two component words.

[4] https://github.com/crazydonkey200/tensorflow-char-rnn

[5] These settings were used for a pre-trained language model that is distributed with this implementation, and so we adopted them as our defaults.

[6] https://github.com/yohasebe/wp2txt

| Dataset | $\text{Comp}_1$ | $\text{Comp}_2$ | Salehi et al. (2015) |
|---|---|---|---|
| ENC | *0.239 | *0.286 | 0.717 |
| EVPC: verb | 0.012 | 0.019 | 0.289 |
| EVPC: particle | *0.313 | *0.301 | - |
| GNC | −0.033 | −0.096 | 0.400 |

Table 2: Pearson's correlation ($r$) for each dataset, using $\text{comp}_1$ and $\text{comp}_2$. Significant correlations ($p < 0.05$) are indicated with *. The best results from Salehi et al. (2015) using $\text{comp}_1$ with representations of the MWE and component words obtained from word2vec (Mikolov et al., 2013), are also shown.

**EVPC** This dataset consists of 160 English verb-particle constructions (e.g., *add up*, *figure out*) which are rated on a binary scale for the compositionality of each of the verb and particle component words (Bannard, 2006) by multiple annotators; no ratings for the overall compositionality of MWEs are provided in this dataset. The binary compositionality judgements are converted to continuous values as in Salehi et al. (2015) by dividing the number of judgements that an expression is compositional by the total number of judgements.

**GNC** This dataset contains 244 German noun compounds (e.g., *Ahornblatt* 'maple leaf', *Knoblauch* 'garlic') which are annotated on a scale of [1,7] for their overall compositionality, and the compositionality of each component word (von der Heide and Borgwaldt, 2009).

### 3.4 Evaluation Methodology

We evaluate our proposed approach following Salehi et al. (2015) by computing Pearson's correlation between the predicted compositionality (i.e., from either $\text{comp}_1$ or $\text{comp}_2$) and human ratings for overall compositionality. For EVPC, no overall compositionality ratings are provided. In this case we report the correlation between the predicted compositionality scores and both the verb and particle compositionality judgements.[7]

## 4 Results

We begin by considering results using the default settings (described in section §3.1) using both $\text{comp}_1$ and $\text{comp}_2$. For $\text{comp}_1$, we set $\alpha$ to 0.7 for ENC and GNC following Salehi et al. (2015); for EVPC we set $\alpha$ to 0.5. Results are shown in table 2. For ENC, and the particle component of EVPC, both $\text{comp}_1$ and $\text{comp}_2$ achieve significant correlations (i.e., $p < 0.05$). However, for GNC, and the verb component of EVPC, neither approach to predicting compositionality gives significant correlations. These correlations are well below those of previous work. For example, using $\text{comp}_1$ with representations of the MWE and component words obtained from word2vec (Mikolov et al., 2013), Salehi et al. (2015) achieve correlations of 0.717, 0.289, and 0.400 for ENC, the verb component of EVPC, and GNC, respectively.[8] Nevertheless, the results in table 2, and in particular the significant correlations for ENC and the particle component of EVPC, indicate that character-level neural network language models do capture some information about the compositionality of MWEs, at least for certain types of expressions.

We now consider the compositionality of individual component words. Because of the low correlations on GNC in the previous experiments, we do not consider it further here. In this case, we compute the compositionality of a specific component word as below, where C is the vector representation of a component word.

$$\text{comp}(\text{C}) = sim(\text{MWE}, \text{C}) \tag{3}$$

Note that this corresponds to $\text{comp}_1$ with $\alpha = 1$ or 0, in the case of the first and second component words, respectively. We compare these compositionality predictions with the human judgements for

[7] In this case Salehi et al. (2015) took the verb compositionality as a proxy for the overall compositionality, and did not consider particle compositionality.

[8] Salehi et al. (2015) did not consider the compositionality of the particle component for EVPC.

| Dataset | word 1 | word 2 |
|---|---|---|
| ENC | 0.135 | *0.335 |
| EVPC | 0.019 | *0.200 |

Table 3: Pearson's correlation ($r$) for the ENC and EVPC datasets for each of component word 1 and 2. Significant correlations ($p < 0.05$) are indicated with *.

| Dataset | | Number of MWEs | $\text{Comp}_1$ | $\text{Comp}_2$ |
|---|---|---|---|---|
| ENC | Attested | 66 | *0.296 | *0.372 |
| | Unattested | 24 | 0.040 | 0.049 |
| EVPC: verb | Attested | 147 | 0.019 | 0.010 |
| | Unattested | 13 | 0.034 | *−0.208 |
| EVPC: particle | Attested | 147 | *0.313 | *0.286 |
| | Unattested | 13 | 0.366 | 0.385 |
| GNC | Attested | 167 | 0.009 | −0.067 |
| | Unattested | 77 | −0.110 | −0.154 |

Table 4: Pearson's correlation ($r$) for MWEs that are attested, and unattested, in each dataset, using $\text{comp}_1$ and $\text{comp}_2$. Significant correlations ($p < 0.05$) are indicated with *. The number of attested and unattested MWEs in each dataset is also shown.

the compositionality of the corresponding component word. Results are shown in table 3. For EVPC, the results are perhaps not surprising given the previous findings, with a significant correlation being achieved for the particle (word 2) but not the verb (word 1). In the case of ENC, a significant correlation is also achieved for the second component word, but not the first.

The above results suggests that the model is better able to predict the compositionality of the second component word of an MWE than the first. To determine whether there is a relationship between the directionality of a character-level language model and the compositionality information it can capture, we also consider a backward LSTM that was trained by reversing the training corpus. The MWE and its component words were then reversed when computing compositionality. However, none of the correlations from this approach were significant.

One interesting aspect of our proposed model is that it can potentially predict the compositionality of out-of-vocabulary expressions that are not observed in the training corpus. In table 4 we present results for each dataset, in the same setup as for table 2, but computing the correlation separately for MWEs that are attested, and unattested, in the training corpus. For ENC, both compositionality measures achieve significant correlations for attested expressions, but not for unattested ones, suggesting that the model cannot predict the compositionality of unseen expressions. In the case of the compositionality of the particle component of EVPC, for both $\text{comp}_1$ and $\text{comp}_2$, the correlations for the unattested expressions are higher than for the attested ones, although for unattested expressions the correlations are not significant. The relatively small number of unattested expressions in EVPC (13) could play a role in this finding. To further investigate this, we focused on expressions in EVPC with less than 5 usages in the training corpus. There are 71 such expressions. For the compositionality of the particle component, $\text{comp}_1$ and $\text{comp}_2$ achieve correlations of 0.327 and 0.308, respectively. These correlations are significant ($p < 0.05$). Word embedding models — such as that used in the approach to predicting compositionality of Salehi et al. (2015) — typically do not learn representations for low frequency items.[9] These results demonstrate that the proposed model is able to predict the compositionality for low frequency items, that would not typically be in-vocabulary for word embedding models, and for which compositionality models based only on word embeddings would not be able to make predictions.[10] For GNC, and the verb component

[9] Salehi et al. (2015) used a minimum frequency of 15, for example.

[10] Note, however, that Salehi et al. (2015) were able to make predictions for all items in EVPC because they trained on a larger corpus (full Wikipedia dumps, as opposed to samples of them) and all items in this dataset were sufficiently frequent in

of EVPC, in line with the previous results over the entire dataset, neither compositionality measure gives significant correlations, with the exception of the verb component of EVPC using $comp_2$ for unattested expressions, although again the number of expressions here is relatively small.

In an effort to improve on the default setup we considered a range of model variations. In particular we considered an RNN and GRU (instead of an LSTM), character embeddings of size 25 and 50 (instead of a one-hot representation), increasing the batch size to 100 (from 20), using dropout between 0.2–0.6, and using a bi-directional LSTM. None of these variations led to consistent improvements over the default setup.

## 5 Conclusions

In this paper we proposed an approach to predicting the compositionality of multiword expressions based on a character-level neural network language model. To the best of our knowledge, this is the first work to consider such character-level models for this task. Our proposed character-level approach has an advantage over prior approaches to compositionality prediction based on distributed representations of words in that we do not require token-level identification of MWEs in order to form representations of them. Our proposed approach can furthermore potentially predict the compositionality of out-of-vocabulary MWEs that are not observed in the training corpus. We carried out experiments over three compositionality datasets: English and German noun compounds, and English verb-particle constructions. Our experimental results indicate that character-level neural network models do capture knowledge of multiword expression compositionality, at least in the case of English noun compounds and the particle component of English verb-particle constructions. We further find that our proposed model captures knowledge of the compositionality of the particle component of English verb-particle constructions that are low frequency or not observed in the training corpus, but not of the compositionality of unobserved English noun compounds.

In future work we intend to further explore the various parameter settings of the language model — such as the batch size, learning rate, and dropout — to better understand their impact on MWE compositionality prediction. We also intend to train the language model on larger corpora. Finally, we intend to combine our character-level approach to compositionality prediction with approaches based on other sources of information, for example distributed representations of words and knowledge from translation dictionaries (Salehi et al., 2014). Specifically, we intend to determine whether the compositionality information from character-level neural network language models is complementary to that in these other approaches.

## References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Association for Computational Linguistics, Sapporo, Japan, pages 89–96.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing* 2:267–292.

Colin James Bannard. 2006. *Acquiring phrasal lexicons from corpora*. Ph.D. thesis, University of Edinburgh.

Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California, pages 242–245.

their training corpus.

Yi Chen, Ming Zhou, and Shilong Wang. 2006. Reranking answers for definitional qa using language modeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1081–1088.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Michael Collins, Brian Roark, and Murat Saraclar. 2005. Discriminative syntactic language modeling for speech recognition. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 507–514.

Leonardo dos Santos Pinheiro and Mark Dras. 2017. Stock market prediction with deep learning: A character-based neural language model for event-based trading. In *Proceedings of the Australasian Language Technology Association Workshop 2017*. pages 6–15.

Waseem Gharbieh, Virendrakumar Bhavsar, and Paul Cook. 2017. Deep learning models for multiword expression identification. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Association for Computational Linguistics, Vancouver, Canada, pages 54–64.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

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

Su Nam Kim and Timothy Baldwin. 2010. How to pick out token instances of english verb-particle constructions. *Language Resources and Evaluation* 44(1–2):97–113.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, pages 636–644.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1520–1530.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at the International Conference on Learning Representations, 2013*. Scottsdale, USA.

Tomáš Mikolov, Ilya Sutskever, Anoop Deoras, Hai-Son Le, Stefan Kombrink, and Jan Cernocky. 2012. Subword language modeling with neural networks. *preprint (http://www. fit. vutbr. cz/imikolov/rnnlm/char. pdf)* .

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science* 34(8):1388–1429.

Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 267–274.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. pages 210–218.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *Proceedings of the 14th Conference of the EACL (EACL 2014)*. Gothenburg, Sweden, pages 472–481.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, pages 977–983.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 194–204.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pages 1818–1826.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association of Computational Linguistics* 2:193–206.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001)*. Hong Kong, China, pages 100–108.

Sabine Schulte im Walde, Stefan Müller, and Stefan Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. Atlanta, GA, pages 255–265.

Raymond Hendy Susanto, Hai Leong Chieu, and Wei Lu. 2016. Learning to capitalize with character-level recurrent neural networks: An empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2090–2095.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu unter-, basis-und oberbegriffen. eine explorative studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*. pages 51–74.