The language-invariant aspect of compounding: Predicting compound meanings across languages

Fritz Günther
University of Milano-Bicocca
Milan, Italy
fritz.guenther@unimib.it

Marco Marelli
University of Milano-Bicocca
Milan, Italy
marco.marelli@unimib.it

Abstract

English. In the present study, we investigated to what extent compounding involves general-level cognitive abilities related to conceptual combination. If that was the case, the compounding mechanism should be largely invariant across different languages. Under this assumption, a compositional model trained on word representations in one language should be able to predict compound meanings in other languages. We investigated this hypothesis by training a word embedding-based compositional model on a set of English compounds, and subsequently applied this model to German and Italian test compounds. The model partially predicted compound meanings in German, but not in Italian.

Italiano. *In questo lavoro abbiamo investigato quanto la composizione sottenda abilità cognitive generali relata alla combinazione concettuale. Se questo fosse il caso, il meccanismo composizionale dovrebbe variare in maniera limitata tra diverse lingue. Di conseguenza, un modello composizionale basato su rappresentazioni lessicali in una data lingua dovrebbe essere in grado di predire significati composizionali in altre lingue. Abbiamo testato questa ipotesi addestrando un modello composizionale sui word embeddings di un set di composti inglesi, e successivamente testato lo stesso modello su composti tedeschi e italiani. Il modello è in grado di predire in modo parzialmente corretto le rappresentazioni dei composti in tedesco, ma non italiano.*

1 Introduction

Compounds are complex words such as *airport*, with two constituents that can be used as free words. Compounding is a highly prevalent phenomenon across many languages. It has been argued to be a proto-linguistic structure to combine simple words into novel and complex concepts, from which more complex compositional language structures have been derived (Jackendoff, 2002).

Given the prevalence and ubiquity of compounding across languages, it is reasonable to assume that speakers of different languages rely, to some degree, on the same cognitive mechanisms to compose the meanings of constituents into a compound meaning. Indeed, the linguistic phenomenon of compounding is generally considered to be the linguistic mirror of the cognitive process of *conceptual combination* (Gagné and Spalding, 2009; Murphy, 2002). Thus, while specific aspects of compounding will inevitably vary between languages due to differences in the language structure and other idiosyncracies, we assume that there is also a language-invariant aspect of compounding that can be transferred across languages. We will investigate this hypothesis by examining whether a compositional model trained on one language (English) is able to predict compound meanings in other languages (German and Italian).

2 Compositional Model

In our study, word meanings are represented via word embeddings derived from large corpora using the *word2vec* model (Mikolov et al., 2013). As a model to derive compound meaning representations from these vectors, we employ the CAOSS model (Marelli et al., 2017), which relies on the compositional model for distributional word vectors proposed by Guevara (2010).

The CAOSS model computes the meaning of a compound as

$$c = M \cdot u + H \cdot v \qquad (1)$$

, where c is the n-dimensional vector representing the compound meaning, u and v are the n-dimensional vectors representing the first and second constituent, respectively, and M and H are $n \times n$-dimensional weight matrices updating the free word meanings into constituent meanings before they are combined.

The weight matrices M and H are estimated through a training procedure on all compound words available in the source corpus for the word embeddings. They are estimated in a least-square regression procedure aimed at optimally predicting these observed compound meanings c from the constituent meanings u and v, following Equation 1.

3 Evaluation Material

In order to investigate our hypothesis, we employed three sets of compounds, collected from various sources: The English set consisted of 5,618 compounds in closed form, collected from the words tagged as noun-noun combinations in the CELEX database (Baayen et al., 1995) and the English Lexicon Project (Balota et al., 2007), and in hyphenated form, collected from the *ukWaC* corpus as described below. The German set consisted of 3,451 compounds in closed form, collected from (Brysbaert et al., 2011) and the GhostNN database (Schulte im Walde et al., 2016). The Italian set of 216 compounds in closed form, collected by one of the authors from an Italian dictionary (Sabatini and Coletti, 2007). Note that the Italian dataset is smaller than the other sets, since compounds are far less common in Italian than in English or German, where compounds are extremely prevalent and compounding is highly productive.

No restrictions based on linguistic criteria (such as endocentric vs. exocentric, or head-first vs. head-second) were applied in the selection of the compounds.

4 Inducing Word Vectors and Training the Compositional Model

4.1 Word Embeddings

Word embeddings were trained on three different web-based corpora (`http://wacky.sslmit.unibo.it`): The English 2 billion word corpus *ukWaC*, the German 1.7 billion word corpus *deWaC*, and the Italian 2 billion word vorpus *itWaC*. While these corpora are not parallel corpora, they were collected using the same web crawler run on different domains (.uk, .de, and .it, respectively). Furthermore, they are very large corpora, which should lead to highly averaged word meaning representations within all three languages. From each of these corpora, *word2vec* word embeddings were derived using the parameter set shown to produce the best results by Baroni et al. (2014): The *cbow* algorithm with a context window size of 5 words producing 400-dimensional vectors (negative sampling with $k = 10$, subsampling with $t = 1e^{-5}$). Word embeddings were only trained for words that occurred more than 50 times in a source corpus.

4.2 Second-Level Vectors

Obviously, the three different semantic spaces were not comparable to one another, as each set of word vectors was trained only on a single-language corpus. Since the weights specified in the matrices M and H of the CAOSS model encode how much each output dimension value for the constituent-updated vectors Mu and Hv is influenced by each input dimension value of the word vectors for the constituents u and v, we could not reasonably apply the CAOSS model trained in one language to word embeddings in another language. We needed word vectors whose dimensions are comparable across the three languages. To this end, we decided to construct *second-level vectors* from the original word embeddings.

The basis for these second-level vectors is the observation that, while word embeddings are not comparable between languages, the similarity *structure* between sets of words is highly comparable across languages. We exploit this observation to define second-level vectors as vectors of similarities between the target and an ordered list of content words (see Table 1). By choosing a list of content words that are as unambiguous as possible and have clear translations across all three languages (such as *pizza*, *Pizza*, *pizza*), we aimed at keeping the second-level vector entries as comparable as possible across languages. We constructed a list containing 300 such aligned content words. With these words, we can demonstrate

original word embeddings				
	dim1	dim2	dim3	...
tomato$_{en}$	0.58	-0.66	-0.92	....
Tomate$_{de}$	-0.23	0.12	0.20	....
pomodoro$_{it}$	-0.01	0.39	-1.37	....
second-level vectors				
en	red	pizza	horse	...
de	rot	Pizza	Pferd	...
it	rosso	pizza	cavallo	...
tomato$_{en}$	0.22	0.28	0.07	....
Tomate$_{de}$	0.23	0.30	0.12	....
pomodoro$_{it}$	0.23	0.26	0.04	....

Table 1: An example for dimensional values of original and second-level word embeddings.

that the similarity structure between words is indeed comparable across languages: We computed all pairwise similarities between these 300 words within each language, and then compared this list of similarities across languages. Similarity correlations across the three languages are substantial: $r = .77$ for English-German, $r = .76$ for English-Italian, and $r = .79$ for German-Italian.

With this aligned list, we converted our word embeddings into second-level vectors by computing, within each language, the cosine similarities between each word in the original semantic space and the 300 content words (see Table 1).

4.3 Evaluation of Second-Level Vectors

In order to serve as adequate word vectors for our compositional model, these second-level vectors need to satisfy two criteria: Firstly, they must adequately capture the similarity structure of the original word embeddings within each language, in order to be used as a substitute for the original word embeddings. Secondly, they have to align word vectors between the three languages: for example, the second-level vector for *tomato* in English should be very similar to the second-level vector for *Tomate* in German and for *pomodoro* in Italian.

Within-Language Reliability. To test for within-language constancy, we first computed the pairwise cosine similarities between all compound constituents from these item sets. Additionally, we computed the cosine similiarities between each compound and its two constituents within each language. These are valid test sets for our study since these are the very embeddings employed to run and test our compositional model later on. In a next step, we computed the same similarities using not the original word embeddings, but the second-level vectors. We then calculated correlations between all the similarity scores computed from the two different vector sets for each of the three languages.

For English, the correlation between the pairwise constituent similarities (2,386 different constituents) was $r = .86$, and the correlation between the constituent-compound similarities was $r = .79$. For German, the correlation between the pairwise constituent similarities (1,929 different constituents) was $r = .80$, and the correlation between the constituent-compound similarities was $r = .72$. For Italian, the correlation between the pairwise constituent similarities (568 different constituents) was $r = .81$, and the correlation between the constituent-compound similarities was $r = .74$. Thus, the similarity structure of the original semantic spaces is to a large extent captured by the second-level vectors, which qualifies them as reliable word meaning representations for our study.

Between-Language Alignment. We tested the across-language alignment of the second-level vectors by means of the original list of 300 content words. This list was constructed to include words that have single clear translation across all three languages. Thus, if the second-level vectors are indeed aligned across the three languages, the three vectors representing these translated words in each language should be very similar to one another.

To test this, we computed the cosine similarity between each of the three translations of these words across the three languages. Using the original word embeddings, the average similarities were virtually zero, as expected for different model trained on different languages: $M = .01$ for English-German, $M = -.00$ for English-Italian, and $M = .01$ for German-Italian. However, computing the same similarities from the second-level vectors improved results dramatically: $M = .80$ for English-German, $M = .80$ for English-Italian, and $M = .82$ for German-Italian. Thus, the second-level vectors are to a large extent aligned across languages, providing the ground to apply a composition model trained on vectors in one language on vectors of the other languages.

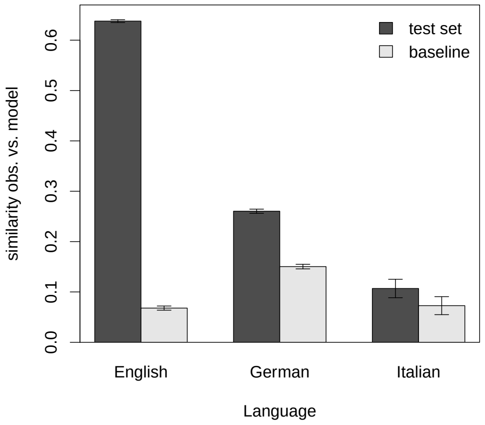

Figure 1: Similarities (mean values and .95 confidence intervals) between observed and model-derived (second-level) vectors for compounds across the three different languages.

4.4 Training the CAOSS model

The CAOSS model was trained on the English second-level word vectors. As a training set, we employed the set of 5,618 English compounds described in the section *Evalutation Material*. The other two languages, German and Italian, were not considered during training.

5 Results

Using the matrices M and H obtained from this training, we computed, for each compound in our evaluation sets, its compound meaning as predicted from the compositional CAOSS model (see Equation 1). The model trained on English was used to compute the model-derived compound meanings for all three languages. We then computed the cosine similarities between these predicted meanings and the corresponding, actually "observed" compound meanings (their respective second-level vectors; e.g. *airport* – [*air*+*port*]). As a baseline comparison level within each language, we computed similarities between the observed compound meanings and model-derived meanings for a random pair of nouns (such as *airport* – [*spring*+*feeling*]). The mean similarities are displayed in Figure 1.

For English, on which our CAOSS model was trained, we obtained a mean similarity between model-derived and observed vectors of $M = .64$, which was significantly above the random baseline ($t(5617) = 122.4, p < .001$).

For the German evaluation set, the mean similarity between model-derived and observed vectors was $M = .26$, which is significantly above baseline ($t(3450) = 20.12, p < .001$).

In contrast, for the Italian evaluation set, the actual similarities did not beat the baseline ($t(215) = 1.39$, $p = .165$). Note that Italian compounds can be classified into head-first compounds (such as *pescespada* – *swordfish*, lit. *fishsword*) or head-second compounds (such as *funivia* – (lit.) *ropeway*)[1]. However, the actual similarities did not beat the baseline in either case ($t(58) = 1.67$, $p = .100$ for head-first compounds; $t(156) = 0.56, p = .578$ for head-second compounds).

The mean value in English differed significantly from German ($t(6460) = 75.53$, $p < .001$), which in turn differed significantly from Italian ($t(238) = 8.18, p < .001$).

6 Discussion

Our results show that a compositional model trained in one language exclusively (English) can be applied to another language (German) to partially predict the meanings of compounds in the latter, of which the model had no training experience at all. Obviously, the model trained on English compounds predicted English compound meanings far better than German compound meanings. This does not stand contrary to our hypothesis: We do not assume that compounding is a tout-court language-invariant mechanism, but that compounding also encompasses general mechanisms *besides* language-specific features.

However, the model trained on English was not able to predict Italian compound meanings above baseline level. Thus, our results only partially support our hypothesis. In interpreting this finding, it has to be considered that the Italian evaluation set was far smaller than the English and the German sets, leading to decreased statistical power in this case (note that, on a purely descriptive level, model performance in Italian is slightly above baseline). Keeping that in mind, our results indicate that the applicability of a compositional model across languages seems to depend on the similarity between the language in which a model was trained and the one where it is applied.

[1] The head is the compound constituent that denotes the semantic category of a word: an *airport* is a type of *port*.

In structural terms, German is in fact much more similar to English than Italian. Both English and German are West-Germanic languages which almost exclusively produce head-second compounds and have highly productive and very rich compounding systems. Italian compounds however can be head-first or head-second, and the compounding system is far less productive in Italian than in English or German (one of the factors responsible for the fact that our Italian item set was smaller than the English or German sets). This explanation is still tentative given the restricted range of languages investigated here. A more thorough investigation on this specific issue would require tests on a wide range of languages, which should be theoretically characterized in terms of their structural similarity with respect to compounding beforehand.

Additionally, future work is required to address other language-dependent aspects of compounding. For example, we focussed only on closed-form compounds, while some languages (for example English and Italian, but not German) can produce open forms such as *school bus* or *pesce spada*. Another issue to be investigated more closely is headedness. On the one hand, head-second Italian compounds are more similar to English and German from a structural point of view; on the other hand, head-first compounds are assumed to be more like English and German in terms of productivity and regularity of meaning. Although our item set included head-first as well as head-second Italian compounds, both are obviously still smaller than the complete Italian item set. Thus, in future studies larger item sets are required to provide such differential tests with the necessary statistical power.

Acknowledgments

This work was supported by Research Fellowship 392225719 from the German Research Foundation (DFG), awarded to Fritz Günther.

References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX lexical data base (CD-ROM).

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39:445–459.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014*, pages 238–247, East Stroudsburg, PA. ACL.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect. *Experimental psychology*, 58:412–424.

Christina L. Gagné and Thomas L. Spalding. 2009. Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, 60:20–35.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, pages 33–37. Association for Computational Linguistics.

Ray Jackendoff. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford, UK.

Marco Marelli, Christina L. Gagné, and Thomas L. Spalding. 2017. Compounding as abstrat operation in semantic space: A data-driven, large-scale model for relational effects in the processing of novel compounds. *Cognition*, 166:207–224.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781v3*.

Gregory L. Murphy, 2002. *Conceptual Combination*, pages 443–475. MIT Press, Cambridge, MA.

Francesco Sabatini and Vittorio Coletti. 2007. *Dizionario della lingua italiana*. RCS Libri, Milano, Italy.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. G$_h$oSt-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2285–2292, Portoroz, Slovenia.