What Meaning-Form Correlation Has to Compose With: A Study of MFC on Artificial and Natural Language

Timothee Mickus[*]
ATILF
tmickus@atilf.fr

Timothée Bernard[*]
Université de Paris
timothee.bernard
@ens-lyon.org

Denis Paperno
Utrecht University
d.paperno@uu.nl

Abstract

Compositionality is a widely discussed property of natural languages, although its exact definition has been elusive. We focus on the proposal that compositionality can be assessed by measuring meaning-form correlation. We analyze meaning-form correlation on three sets of languages: (i) artificial toy languages tailored to be compositional, (ii) a set of English dictionary definitions, and (iii) a set of English sentences drawn from literature. We find that linguistic phenomena such as synonymy and ungrounded stop-words weigh on MFC measurements, and that straightforward methods to mitigate their effects have widely varying results depending on the dataset they are applied to. Data and code are made publicly available.

1 Introduction

Compositionality is one of the core aspects of language: when speaking, we weave complex propositions out of unanalyzable atoms of meaning. This intricate interplay between symbols and sentences has been an important point of research in areas ranging from philosophy of language (Frege, 1884) to distributional semantics (Baroni et al., 2014), and has been held as one of the key characteristic differentiating human language from animal communication (Hockett, 1960). Compositionality in natural language is usually captured by the Principle of Compositionality. This principle, often attributed to Frege although this paternity is debatable (Janssen, 2012), has been worded in different ways; let us quote a classic textbook (Gamut, 1991, p.26): "the meaning of a composite expression must be wholly determined by the meanings of its composite parts and of the syntactic rule by means of which it is formed".

One suggestion, that can be traced back to Kirby (1999), and that was fully operationalized in Kirby et al. (2008), is that compositionality can be measured as a correlation between meaning and surface form (i.e., the sequence of tokens). If we consider sentences such as "*I saw a black cat*" and "*I saw a black crow*", we see that minute changes in the form of a sentence entail that its meaning does not change too much. Compare these examples to sentences such as "*I saw a white crow*", or "*He saw a white crow*": the gradual alteration of the form of a sentence also gradually alters its meaning. Provided that we can measure similarity between any two sentences forms, and between any two sentences meanings, this observation can be rephrased as follows: the distance between two sentences meanings should correlate with the distance between their forms. We refer to this as *meaning-form correlation*, or MFC for short.

One area where MFC has been employed as a way to both detect and quantify compositionality is the field of emergent communication (Kirby et al., 2008; Kirby et al., 2015; Spike, 2016; Ren et al., 2020, a.o.) which studies agents (artificial or human) who have to produce messages in order to express well defined meanings. To estimate the compositionality of a set of message-meaning pairs, one can compute a Pearson or Spearman correlation score between textual distances and meaning distances. Additionally, a permutation test (over random permutation of the message-meaning assignment) can produce a p-value indicating whether the correlation is statistically significant as well as a z-score that quantifies in number of standard deviations the difference between the correlation found in the language and the correlation of

---

[*]Work done while at AIST (Japan).

random assignments. This is done among others by Kirby et al. (2015), Gutiérrez et al. (2016), or Spike (2016) using a Mantel test (Mantel, 1967), which is based on Pearson's notion of correlation.

To the extent of our knowledge, no study has investigated the validity of this methodology. Is MFC a coherent method of measuring compositionality, especially when it comes to *natural language*? Or do other factors weigh on the measurements and obfuscate the results? If so, can we identify these factors, and can we control for them? To address this gap, we therefore study how MFC behaves, both in a controlled setup based on artificial data and in more realistic scenarios involving natural language.[1]

2 MFC and Artificial Languages

MFC-based assessments of compositionality implicitly assume that any change in form should correspond to some change in meaning. This can however be challenged: for instance, synonyms and paraphrases will introduce changes in form that should not entail change in meaning. Thus we expect MFC to be sensitive to such phenomena, and this in turn suggests that factors such as synonymy could overpower the effect of compositionality that we wish to detect using MFC. To approach this question, we generate artificial languages containing varying degrees of compositionality as well as potential confounding factors.

2.1 Methodology

Our experimental protocol consists in generating artificial languages with varying properties, and see how these properties impact MFC measurements. All of our artificial languages are sets of message-meaning pairs. We represent meanings as binary vectors of five components, whereas messages are sequences of symbols. We refer to each of the five semantic dimensions as a *concept*. In most cases, the value of a concept will be denoted in a message by a specific symbol, which we call its *expression*.

As we are interested in whether MFC accurately captures *compositionality*, we design languages where some of the concepts are systematically expressed conjointly by unanalyzable holistic expressions. We generate languages where the values of the first h concepts are systematically expressed through a single expression, and the other $5 - h$ are left untouched, with h varying from 1 to 5. When $h = 1$, the language is entirely compositional; when $h = 5$, the language is entirely holistic.

We also consider *synonymy* as a possible confounding factor: as previously noted, synonyms entail a variation in form that is not coupled with a variation in meaning. To model this phenomenon, we generate languages in which any single value of a concept can equally be expressed by s different expressions, with s ranging from 1 to 3. As a consequence, when $s = 1$, the language exhibits no synonymy.

Moreover, we expect that MFC measurements might be impacted by the presence of *semantically ungrounded elements*—i.e., elements not associated with any concept or combination of concepts. We therefore generate languages where u specific ungrounded symbols appear once in every message at randomly chosen positions, with u varying between 0 and 3. In languages where $u = 0$, the language contains only semantically grounded expressions.

Finally, we consider the case of *paraphrases*, sentences of different forms but equivalent meanings. For a language to contain paraphrases, it must be able to express a single meaning with different messages. This is the case in our artificial languages that exhibit synonymy or contain semantically ungrounded elements: the variation they introduce allow distinct messages to have the same semantic. We thus generate languages for which p messages are produced for each meaning before dropping possible meaning-message pair duplicates. p ranges from 1 to 3. If $p = 1$, the language contains no paraphrase.

We test all possible combinations of these four parameters. We also include *random baselines* where we assign meanings to random sequences of symbols, either of an arbitrarily fixed length of 5 symbols, or of a length chosen uniformly between 1 and 10. We generate 50 artificial languages for every combination of parameters to help us distinguish the stable effects of our parameters from spurious accidents due to our random generation process. We refer to each of the 50 generation processes as a separate *run*. We compute Mantel tests using the Hamming distance between meaning vectors—i.e., the number of differing components—and the Levenshtein distance over messages, normalized by the maximum length

---

[1]Data and code are publicly available at `https://github.com/TimotheeMickus/mf-correl`.
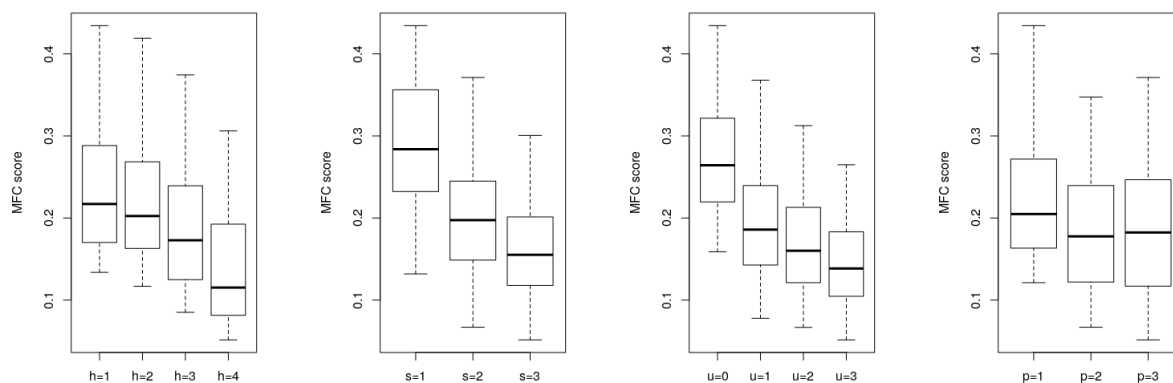
(a) Compositionality	(b) Synonymy	(c) Ungrounded elements	(d) Paraphrase

NB: No $h = 5$ is found significant.

Figure 1: MFC for artificial languages, grouped by parameter. (significant items only, avg. of 50 runs)

of the two messages. For each language, we study the corresponding p-value and the correlation score. For every combination of parameters, we study its average p-value and correlation score across all runs.

One limitation of this methodology is that our modeling may not comply fully with natural language—in particular, the existence of exact synonyms is debatable; likewise, natural function words do possess some semantic content, whereas our ungrounded symbols do not. Neither do we claim to conduct an exhaustive study of all relevant phenomena: we leave other likely factors such as multi-word expressions to future work.

2.2 Results

A visualization of the results for the compositional variation factors is shown in Figure 1. Each subfigure corresponds to a different factor, and shows the distribution of MFC scores according to the possible levels for that factor. As expected, random baselines were found to be insignificant (p-value ≥ 0.05).

If we focus on compositionality (Figure 1a), we do see that less compositional languages yield lower MFC scores. When we consider the correlation values averaged over all 50 runs, we see that no holistic parameter configuration (where $h = 5$) is found to be significant, resulting in the missing boxplot in Figure 1a. 1st, 2nd and 3rd quartiles are found to consistently decrease for higher values of h.[2]

Synonymy and semantically ungrounded elements are found to be confounding factors (Figures 1b and 1c). Higher values for the s and u parameters systematically entail that the distribution of MFC scores (averaged over 50 runs) is globally lower, as 1st, 2nd and 3rd quartiles consistently decrease.

Lastly, for non fully-holistic settings ($h < 5$), we observe that some combinations of factors fail to produce significant MFC scores. In 15.2% of all possible combinations, as this persists even when averaged over all runs, we conclude that this is an actual effect of the interaction of factors. All these languages are defined with at least one extreme factor: viz. either three synonyms per concept ($s = 3$), three ungrounded symbols ($u = 3$) or four concepts merged into a single expression ($h = 4$); moreover all of them (except for two languages defined with $h = 4$, $s = 3$ and either $u = 2$ or $u = 3$) contained a single message per meaning ($p = 1$); confirming this trend, we find that all non fully-holistic languages with up to three messages per meaning ($p = 3$) were found to have a significant MFC on average. These shared characteristics can hint at the fact that confounding factors can significantly obfuscate the compositional structure of the messages. An alternative explanation could be that languages with-

---

[2]Some languages with $h = 5$, which are fully holistic, were found to yield significant MFC. Most of these also included multiple ungrounded symbols ($u \geq 1$). This can be explained by the effects of paraphrases: in holistic languages with multiple messages per meaning containing ungrounded symbols, paraphrastic messages for a given meaning differ only by their ungrounded symbols, whereas messages for different meanings will also differ by their grounded symbols—leading to nonzero correlation. However, on average over all 50 runs, the p-value for any of these settings is below our threshold.

Coeffs.	Estimate	Std. Error	t value	Pr(> $|t|$)
(Intercept)	0.402 781	0.001 936	208.07	$< 2 \cdot 10^{-16}$
$h = 2$	−0.016 909	0.001 577	−10.72	$< 2 \cdot 10^{-16}$
$h = 3$	−0.057 501	0.001 577	−36.46	$< 2 \cdot 10^{-16}$
$h = 4$	−0.113 547	0.001 577	−72.00	$< 2 \cdot 10^{-16}$
$h = 5$	−0.197 191	0.001 734	−113.75	$< 2 \cdot 10^{-16}$
$s = 2$	−0.096 726	0.001 748	−55.34	$< 2 \cdot 10^{-16}$
$s = 3$	−0.137 808	0.001 748	−78.85	$< 2 \cdot 10^{-16}$
$u = 1$	−0.099 435	0.001 524	−65.26	$< 2 \cdot 10^{-16}$
$u = 2$	−0.126 905	0.001 524	−83.29	$< 2 \cdot 10^{-16}$
$u = 3$	−0.145 707	0.001 524	−95.63	$< 2 \cdot 10^{-16}$
$p = 2$	0.031 989	0.001 354	23.63	$< 2 \cdot 10^{-16}$
$p = 3$	0.041 219	0.001 354	30.45	$< 2 \cdot 10^{-16}$

Table 1: Linear model of correlation with parameters as predictors. Intercept: $h = 1, s = 1, u = 0, p = 1$.

out paraphrases ($p = 1$) contain fewer messages and thus yield higher p-values, whereas paraphrases additionally entail that very low textual distances map to zero meaning distances.

2.3 Discussion & Conclusions

We observed that synonymy (Figure 1b) and ungrounded elements (Figure 1c) seemed detrimental to MFC scores, whereas the effects of paraphrases were found to be more subtle (Figure 1d). We quantify this by computing a simple linear model in R (R Core Team, 2018) where the correlation score is the dependent variable and the values of the four parameters are the predictors; data points correspond to specific runs. Results are reported in Table 1. While $h = 5$ was found to be the predictor with the strongest negative effect on MFC scores, we found that factors $s = 3$, $u = 3$ and $u = 2$ had stronger effects than $h = 4$. In short, the model shows that factors such as synonymy impact MFC measurements—sometimes to a greater extent than compositionality as shown by t-value scores. It also stresses that paraphrases positively impact MFC scores: in languages where $p > 1$, any single unreliable message is less likely to whittle down scores.

In all, our experiment suggests that taking MFC as an indicator for compositionality comes with significant challenges. Factors that we expect from natural language, such as ungrounded symbols and synonyms, obfuscate the clear relationship between compositionality and MFC scores. At times, these factors can even annihilate MFC scores of compositionally generated languages. Yet compositionality does impact measurements: therefore, while MFC scores in and of themselves may not be sufficient to establish or reject the compositionality of a given set of messages, they can serve as a diagnosis tool.

3 MFC and Natural Language: Definitions

The experiments we conducted in Section 2 have revealed that confounding factors can impact MFC scores on artificial languages. This raises the question of how MFC behaves on *natural* instances of compositionality: can the same confounding variables be observed? Is it possible to control for them? In short, we now aim to translate our results derived from artificial data to natural language data.

Computing Mantel tests for natural examples of composition requires us to select a set of natural language expressions for which we can compute semantic distances between any pair. Taking inspiration from Hill et al. (2016), we use dictionary definitions. Indeed, in dictionaries, the meaning of the word being defined (or "definiendum"; pl. "definienda") is arguably also the meaning of the definition gloss. Hence, we can equate the semantic distance between two glosses to the semantic distance between the corresponding definienda. Distributional semantics models (Lenci, 2018) allow us to conveniently compute the latter: they represent words as vectors in a space over which we can use metrics such as cosine similarity or Euclidean distance, and have moreover been found to match human intuitions (Mandera et al., 2017).
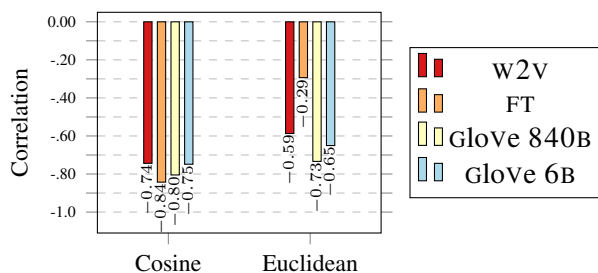
Figure 2: Meaning distance metrics evaluated on the MEN dataset. (Spearman correlation)

3.1 Methodology

We select definitions from the dataset distributed by Noraset et al. (2017). We restrict ourselves to definitions of nouns collected from the GCIDE, where the definiendum is among the 100–10 000 most frequent words of the English Gutenberg corpus.[3] This yields 4 123 distinct definienda and 20 109 definitions. We repeat this process five times for all subsequent measurements before averaging results.

One could arguably compute MFC on the basis of human-annotated semantic similarity ratings. Datasets of such ratings exist (Marelli et al., 2014; Cer et al., 2017, e.g.), but they do not provide a (dis)similarity between every pair of items as is required to perform a Mantel test. Instead, as mentioned above, we use distributional semantics models as proxy for meaning representations and use a distance between word vectors. We consider four sets of pre-trained word embeddings: fastText (Bojanowski et al., 2017), GloVe 6B and GloVe 840B (Pennington et al., 2014), and word2vec (Mikolov et al., 2013).

We first verify that the semantic distances over these semantic spaces properly anti-correlate to human similarity ratings: words that humans judge to be highly similar in meaning should not be far from one another in the embedding spaces. We therefore assess the four sets of word embeddings on a word semantic similarity benchmark, the MEN dataset (Bruni et al., 2014); for all models, we test the Euclidean distance and the cosine distance as a distance over vectors. Results in Figure 2 highlight that while all semantic metrics properly anti-correlate, cosine distance yields higher correlations with human similarity judgments than Euclidean distance, for all embedding spaces.

While it is most standard to build on cosine (the most common semantic vector similarity metric) and Levenshtein distance (consistently with artificial language studies), we also consider alternative setups with Euclidean distance, tree-based form metrics and distance normalization. Unlike the artificial languages we devised earlier, natural language has a rich syntax. It therefore makes sense to assess the textual similarity of these sentences using syntactically-informed metrics. Thus, in addition to the Levenshtein distance over words, we study the Tree Edit Distance (TED) computed with the AP-TED algorithm (Pawlik and Augsten, 2015) over the corresponding parse trees obtained with the Reconciled Span Parser (Joshi et al., 2018). As textual distances should arguably not be sensitive to the size of the sentences, we also consider normalizing them so that the maximum distance between any two definitions is 1.[4]

To check whether synonymy, ungrounded symbols and paraphrases also impact natural language examples, we perform simple modifications of our original process. To control for synonymy, we replace every word by the first lemma of its first synset in WordNet (Fellbaum, 1998), if any such lemma can be found. To control for ungrounded symbols, we remove stop-words from our definitions; as this results in definitions that greatly differ from the parser's training examples, we do not compute TED-based MFC scores in this case. To control for paraphrases, we redo the selection process, and this time randomly sample 4 123 items out of the total 20 109 definitions, without the constraint of having only one definition per definiendum: hence these samples only contain on average 2 507 distinct definienda (standard deviation: ±14.53). We duly note that different definitions correspond to distinct senses of the definiendum and thus are not strictly speaking paraphrases. However, Arora et al. (2018) show that the embedding

---

[3]The GNU Collaborative International Dictionary of English (GCIDE): https://gcide.gnu.org.ua/.

[4]We normalize Levenshtein distance by the length of the longest of the two items. We normalize TED between $\mathcal{F}$ and $\mathcal{G}$ by $\#\mathcal{F} + \#\mathcal{G} - \min\left(h(\mathcal{F}), h(\mathcal{G})\right)$, where $\#\mathcal{T}$ corresponds to the size of tree $\mathcal{T}$ and $h(\mathcal{T})$ to its height.
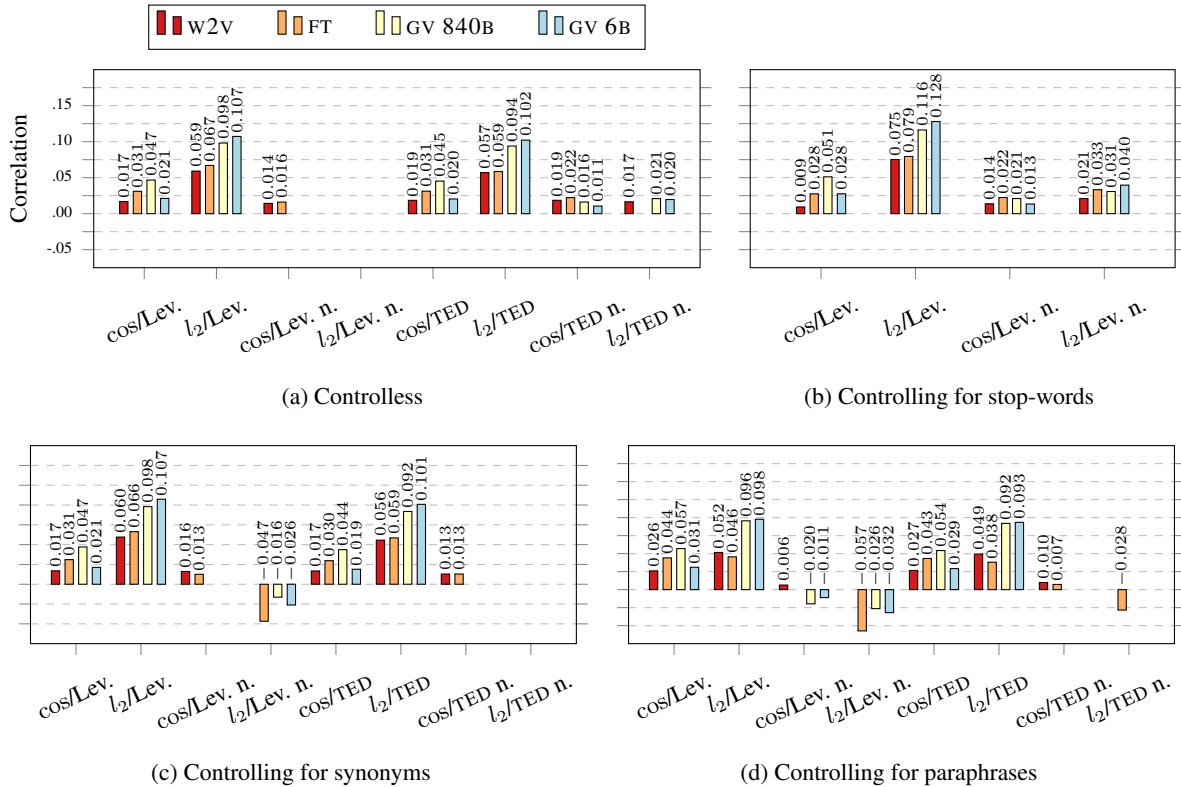
Figure 3: MFC scores for natural language definitions.

of a polysemous word corresponds to a weighted sum of underlying meaning embeddings; thus a word embedding ought to be matched to the entire set of definitions for the corresponding token. In any event, we expect these methods to indicate a consistent trend in MFC scores, despite their crudeness.

3.2 Results

We summarize our findings in Figure 3. Figure 3a corresponds to results in the case where no supplementary control is applied. Figures 3b, 3c and 3d highlight the effects of controlling for stop-words, synonyms, and multiple definitions respectively. Statistically insignificant MFC scores are not displayed. Standard deviations of MFC scores across the five runs remain below 0.008 and often around 0.005, with the exception of a few setups when controlling for paraphrases.[5]

While our crude control methods often produce inconsistent measures, they do suggest that the identified MFC confounds are noticeably at play in natural language.

As with artificial language experiments above, controlling for *stop-words* improved MFC measurements in most setups (Figure 3b), yet scores decrease for fastText and word2vec when using cosine and Levenshtein distances setups. Controlling for *synonymy* brought very small suggestive MFC increases in the most standard setup (cosine and non-normalized Levenshtein) but observations across other setups are not consistent (Figure 3c). Controlling for *paraphrases* produces consistent and pronounced MFC improvements on the most standard setup but quite diverse effects elsewhere (Figure 3d).

We also make two general unexpected observations about alternative setups. First, as seen e.g., from the controlless scenario in Figure 3a, normalizing textual metrics can be surprisingly detrimental. For instance, Euclidean distance between GloVe 6B vectors, when paired with Levenshtein distance, goes from the highest measured MFC score to statistical insignificance—in fact, only 2 out of the 8 MFC using normalized Levenshtein distance are significant. When controlling for confounding variables, normalization can even induce anti-correlations. Second, cosine distance also yields lower MFC scores than Euclidean distance, despite it being more in line with human ratings (Figure 2): Euclidean distance

---

[5]In detail, 4 setups yield a standard deviation between 0.01 and 0.015; they all involve Euclidean distance and non-normalized form distances, using either fastText or GloVe 840B.

yields in many occasions correlation scores of 0.1 and higher, more than twice what we observe for cosine distance.

3.3 Discussion & Conclusions

Overall, this experiment highlights that demonstrating the role of confounding factors that we expect to be detrimental to MFC, such as ungrounded elements, is in principle possible for natural language; though we have employed blunt methods of control, effects could be perceived. On the other hand, our observations underscore how sensitive MFC is to the choice of distance functions: considerations such as normalizing metrics between 0 and 1 or choosing Euclidean vs. cosine distance can impact results significantly. These observations raise two questions: (i) why does normalizing textual distances degrade MFC scores? and (ii) do these results support that MFC captures compositionality in natural language? To answer both, we study more closely which items are detrimental to MFC, as they may shed light on what MFC measurements capture, and how normalization affects it. We consider items where measurements are mismatched: sentences with a relatively low meaning distance but a relatively high form distance—or vice versa—drive the MFC score down. We convert distance measurements into rank values, and consider the 100 pairs of sentences that yield maximal rank difference for a given setup. These pairs will be referred to as *problematic* below.

For the sake of clarity, we focus on GloVe 6B-based measurements, using Levenshtein distance (either raw or normalized) and compute all our measurements from the same random selection of $4\,123$ definitions. In the controlless scenario, the 100 most problematic items for non-normalized Levenshtein-based setups all involved two synonym-based definitions (e.g., "*pilot: a steersman*"). As synonym-based definitions can be identified to holistic messages, this evaluation suggests that our dataset contains a high number of non-compositional examples that this textual metric is not fit to handle. That these items are reliably deemed problematic is coherent with the hypothesis that MFC captures compositionality.

All problematic pairs were found to have a high semantic distance (involving two unrelated definienda) but a low textual distance (both definitions are very short, usually composed of an article and a noun, which entails few edit operations). This suggests that normalizing the textual distance might be crucial to get reliable MFC. However, while normalizing Levenshtein distance does reduce the number of pairs of synonymy-based definitions in the problematic pairs, they remain very frequent ($98/100$ for Euclidean distance, $88/100$ for cosine distance), due to the fact that a common article usually means that half of the tokens in the two definitions are the same. Removing stop-words further reduces synonymy-based definitions in problematic pairs to a handful, but reveals other artifacts: many pairs share a common pattern such as "*sneer: the act of sneering*"/"*wade: the act of wading*". Moreover, the MCF obtained in these settings are lower than with raw Levenshtein distance, suggesting that such patterns are frequent and that length is (counter-intuitively) a relevant factor. We conjecture that short definitions are responsible for a large part of these artifacts, in which case a solution would be to filter them out. Alternatively, a more complex textual distance could allow us to get rid of these artifacts. We leave this subject for future work.

In summary, this experiment suggests that MFC can detect compositionality in natural language, but that the exact setup employed is crucial. Distance normalization appears to annihilate MFC scores and the effects of controlling for confounding factors vary from distances to distances.

4 MFC and Natural Language: Sentence Encoders

The results in Section 3 suggest that MFC scores on natural language reflect its compositionality but also confounding structural factors that obfuscate semantic transparency. However, the findings also raise questions about the peculiarity of the dataset, since dictionary definitions do not form a representative sample of language use by any standards. Is it possible to replicate the methodology on more naturalistic text data? The main difficulty lies in finding a relevant and practical notion of semantic distance for arbitrary sentences, since human judgments can prove arduous to collect. As an attractive potential strategy, one could use vector distances over semantic representations produced by models of semantic composition called sentence encoders. Do these models provide good enough approximations for sentence meaning similarity? If so, we expect to be able to reproduce the effects of confounding factors
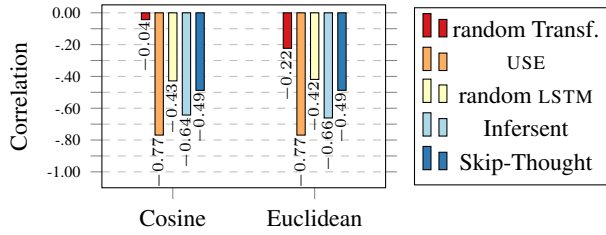
Figure 4: Meaning distance metrics evaluated on the SICK dataset. (Spearman correlation)

(a) No control

(b) Controlling for stop-words
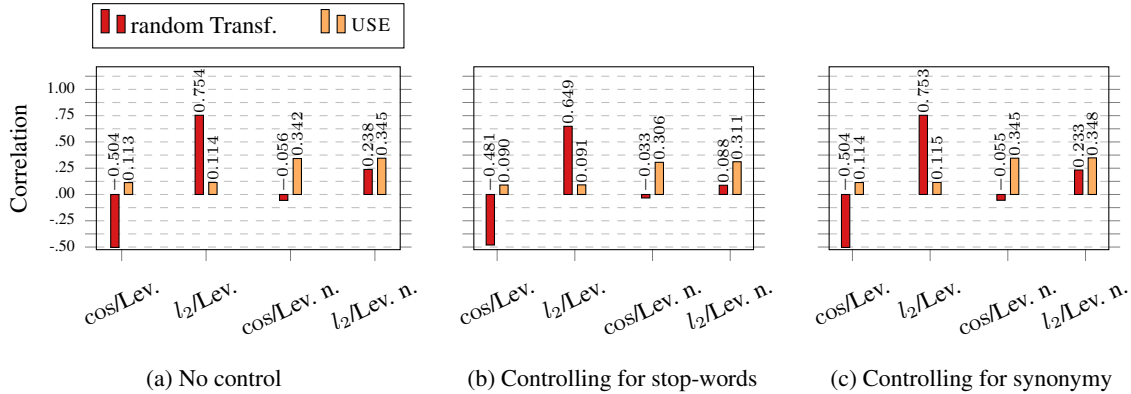
(c) Controlling for synonymy

Figure 5: MFC scores for natural language sentences.

on MFC as predicted by representations derived from sentence encoders. To answer these questions, we replicate our previous experiment on a set of common sentences, rather than definitions.

4.1 Methodology

The semantic distances used in this section rely on *sentence encoders*, computational models that convert sequences of tokens into vector representations. They can be trained on a variety of tasks, from predicting the entailment relation between a pair of sentences (Conneau et al., 2017) to reconstructing the context of a passage (Kiros et al., 2015). These tasks require capturing the meaning of the corresponding texts.

We first assess whether sentence encoders yield coherent representations of meaning by computing the Spearman correlation between the human ratings present in the SICK benchmark (Marelli et al., 2014) and the cosine and Euclidean distances between the two corresponding sentence embeddings; as with word embedding spaces in Section 3, we expect a significant anti-correlation between the two. Figure 4 summarizes correlation scores for Skip-Thought (Kiros et al., 2015), Infersent (Conneau et al., 2017) and the Universal Sentence Encoder (Cer et al., 2018, USE for short), along with a randomly initialized (and untrained) Transformer (Vaswani et al., 2017) for perspective. We observe that USE yields the most consistent semantic representations and thus decide to focus in the following on this particular model.

We randomly sample 4 123 sentences from the Toronto BookCorpus (Zhu et al., 2015)—a collection of English books of various genres—for computing Mantel tests using Levenshtein distance, both raw and normalized, as the textual distance. We repeat the procedure 5 times before averaging results. We use the same operations as previously to control for synonyms and stop-words; i.e., we test whether removing stop-words altogether and normalizing words based on their WordNet synsets improve MFC.

4.2 Results

Results are presented in Figure 5: Figure 5a corresponds to the controlless scenario; Figures 5b and 5c present the effects of controlling for stop-words and synonyms respectively. Results are consistent across all five random samples of sentences: standard deviation of MFC scores is systematically below 0.016,

and often below 0.005.[6]

Most striking are the very high correlations and anti-correlations that are yielded by the random baseline: the anti-correlations and correlations derived from non-normalized Levenshtein distance have a greater magnitude than what we observe for USE (which is also based on the Transformer architecture). In the case of Euclidean distance, this magnitude can partly be explained by the architecture itself, which computes a vector that is not meaningful but that is still computed in a very compositional way, although the corresponding notion of composition is not linguistically justified. The simple sum used to derive the sentence embeddings from the hidden states at each timestep[7] entails that the norm of every sentence representation grows proportionally to the number of words it contains; moreover the residual connections used in Transformers entail that the hidden state for a given timestep bears some trace of the input word at this timestep, thus sentences with words in common will tend to be nearer in the Euclidean space.

Turning to the MFC for USE, we observe that normalizing Levenshtein distance leads to higher scores—opposite to what we found for definitions. On the other hand, cosine-based setups overall are found to decrease scores by a low margin, of 0.005 at most—which is consistent with what was observed for definitions, but much less pronounced. Removing stop-words lowers the correlation for USE embeddings, which is also contrary to what we observed for definitions: scores in Figure 5b are found to be lower than those without any form of control by a margin ranging from 0.02 to 0.04. Lastly, while we technically observe a higher MFC when controlling for synonyms (Figure 5c), the effect is again very subtle: correlation increases by no more than 0.005.

4.3 Discussion & Conclusions

This last experiment is first a reminder of the fact that the use of sentence encoders to compute MFC only makes sense under the assumption that they accurately represent the semantic of their input. While this assumption is obviously not satisfied for a random model, it is defensible for USE given its training procedure and the high anti-correlation observed above in Figure 4.

Second, if we accept the relevance of using this latter model, the results for USE highlight the fact that the issues we encountered with the definition dataset of Section 3, in particular related to distance normalization, are not systematically present. This suggests that the counter-intuitive behavior observed on definitions might be due to particular artifacts from this dataset, in contrast to the more natural and varied dataset used in the present section.

5 Related Works

The core of this work draws upon previous research on MFC (Kirby et al., 2008; Kirby et al., 2015; Spike, 2016; Ren et al., 2020, a.o.). This concept finds its roots in works such as the ones of Kirby (1999), Kirby (2001) or Brighton and Kirby (2006), which suggest that while compositionality in language relies on syntactic structures, it could still be measured using surface form. Other implementations of MFC include studies centered on correlations between form and meaning at the word or sub-morphemic level—conflicting with the assumption of arbitrariness of the sign—such as the work of Gutiérrez et al. (2016), Kutuzov (2017), Dautriche et al. (2017) or Pimentel et al. (2019).

More broadly, compositionality has proven to be a fruitful field of research; see for example the overview edited by Hinzen et al. (2012). The NLP community has produced models of compositional semantics (Baroni et al., 2014; Marelli and Baroni, 2015; Conneau et al., 2017) and significant efforts have been made to assess how well they capture human intuitions (Dinu and Baroni, 2014; Marelli et al., 2014; Cer et al., 2017; Wang et al., 2018, a.o.) and to analyze their behavior (Liska et al., 2018; Baroni, 2020, e.g.). It is worth stressing that assessing the degree of compositionality within a corpus or a model, as do these works, is a distinct enterprise from quantifying the degree of similarity between two sentences (Papineni et al., 2002; Lin, 2004; Clark et al., 2019, e.g.)—more precisely, the methodology of MFC employs the latter to estimate the former.

---

[6]Only the randomly initialized Transformer when using Euclidean distance and non-normalized Levenshtein distance yields standard deviations above 0.01 (in all three scenarios). Standard deviations for USE embeddings are all below 0.003, with the exception of the two setups involving normalized Levenshtein distance and controlling for stop-words (around 0.007).

[7]Like Cer et al. (2018), we divided the sentence embedding by the square root of the length of sentence.

Lastly, another related topic of research concerns the use of dictionaries as meaning inventories in NLP. For instance, Chodorow et al. (1985), Gaume et al. (2004), Tissier et al. (2017) and Bosc and Vincent (2018), all employ dictionary entries for purposes ranging from word-sense disambiguation to ontology population and to embedding computation. Directly relevant here is the work of Hill et al. (2016), who leverage the compositional aspect of definitions to compute sentence meaning representations.

6 Conclusions

In this work, we have empirically assessed the validity of meaning-form correlation as a methodology for studying compositionality, both in artificial and natural languages. In all, the relationship between meaning-form correlation and compositionality is not straightforward; experimental setups that would employ the former to represent the latter ought to proceed with caution. Confounding factors that we can expect from natural language can also weigh on meaning-form correlation in artificial languages, and their effects are not straightforward to control for. Other issues such as the architecture employed to derive meaning representations weigh on the relevance or interpretation of correlation scores. Meaning-form correlation is very dependent on the actual metrics and dataset under consideration. On the other hand, meaning-form correlation was also shown to be able to detect compositionality, both in controlled artificial setups and with natural language data, suggesting that this methodology can be used to study compositionality as long as due attention is paid to confounding factors.

While we have focused on modeling a small number of confounding factors, others can obviously be tested using our methodology. For instance, preliminary experiments suggest that factors such as duality of patterning—i.e., using non-atomic expressions (Hockett, 1960)—reinforce correlations to some extent, whereas aspects such as free ordering of constituents seem detrimental. This limited study has also focused on a specific implementation of meaning-form correlation, based on Mantel tests—however other related setups exist, for example based on mutual information (Pimentel et al., 2019). We leave them to future studies. Another limitation of the present work is that we have focused solely on English: we intend in future experiments to experiment with typologically diverse languages, ranging the full isolating-synthetic spectrum.

Acknowledgements

References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 375(1791), February.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium, October-November. Association for Computational Linguistics.

Henry Brighton and Simon Kirby. 2006. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artif. Life*, 12(2):229–242, March.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47, January.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November. Association for Computational Linguistics.

Martin S. Chodorow, Roy J. Byrd, and George E. Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *23rd Annual Meeting of the Association for Computational Linguistics*, pages 299–304, Chicago, Illinois, USA, July. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy, July. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

Isabelle Dautriche, Kyle Mahowald, Edward Gibson, and Steven T. Piantadosi. 2017. Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8):2149–2169.

Georgiana Dinu and Marco Baroni. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 624–633, Baltimore, Maryland, June. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Gottlob Frege. 1884. *Grundlagen der Arithmetik: Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Wilhelm Koebner, Breslau.

L. T. F. Gamut. 1991. *Logic, Language, and Meaning*, volume 1: Introduction to Logic. University of Chicago Press.

Bruno Gaume, Nabil Hathout, and Philippe Muller. 2004. Word sense disambiguation using a dictionary for sense similarity measure. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

E. Dario Gutiérrez, Roger Levy, and Benjamin Bergen. 2016. Finding non-arbitrary form-meaning systematicity using string-metric learning for kernel regression. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2379–2388, Berlin, Germany, August. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Wolfram Hinzen, Edouard Machery, and Markus Werning. 2012. *The Oxford Handbook of Compositionality*. Oxford University Press, 09.

Charles F. Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–96, September.

Theo Janssen. 2012. Compositionality: Its historic context. In *The Oxford Handbook of Compositionality*, pages 19–46. Oxford University Press, 09.

Vidur Joshi, Matthew Peters, and Mark Hopkins. 2018. Extending a parser to distant domains using a few dozen partially annotated examples. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1190–1199, Melbourne, Australia, July. Association for Computational Linguistics.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Simon Kirby. 1999. Syntax out of learning: The cultural evolution of structured communication in a population of induction algorithms. In *Proceedings of the 5th European Conference on Advances in Artificial Life*, ECAL '99, page 694–703, Berlin, Heidelberg. Springer-Verlag.

Simon Kirby. 2001. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *Trans. Evol. Comp*, 5(2):102–110, April.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Andrei Kutuzov. 2017. Arbitrariness of linguistic sign questioned: Correlation between word form and meaning in russian.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Adam Liska, Germán Kruszewski, and Marco Baroni. 2018. Memorize or generalize? searching for a compositional RNN in a haystack. *AEGAP Workshop, ICML*.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78, 02.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220.

Marco Marelli and Marco Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122 3:485–515.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. 05.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01.

Thanapon Noraset, Chen Liang, Lawrence Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *AAAI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Mateusz Pawlik and Nikolaus Augsten. 2015. Efficient computation of the tree edit distance. *ACM Trans. Database Syst.*, 40(1), March.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy, July. Association for Computational Linguistics.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. 2020. Compositional languages emerge in a neural iterated learning model. In *International Conference on Learning Representations*.

Matthew John Spike. 2016. *Minimal requirements for the cultural evolution of language*. Ph.D. Thesis, University of Edinburgh, Edinburgh, Scotland, UK.

Julien Tissier, Christophe Gravier, and Amaury Habrard. 2017. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.