Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation

Verna Dankers[1], **Christopher G. Lucas**[1], and **Ivan Titov**[1,2]
[1]ILCC, University of Edinburgh
[2]ILLC, University of Amsterdam
vernadankers@gmail.com, {clucas2, ititov}@inf.ed.ac.uk

Abstract

Unlike literal expressions, idioms' meanings do not directly follow from their parts, posing a challenge for *neural machine translation* (NMT). NMT models are often unable to translate idioms accurately and over-generate compositional, literal translations. In this work, we investigate whether the non-compositionality of idioms is reflected in the mechanics of the dominant NMT model, Transformer, by analysing the hidden states and attention patterns for models with English as source language and one of seven European languages as target language. When Transformer emits a non-literal translation – i.e. identifies the expression as idiomatic – the encoder processes idioms more strongly as single lexical units compared to literal expressions. This manifests in idioms' parts being grouped through attention and in reduced interaction between idioms and their context. In the decoder's cross-attention, figurative inputs result in reduced attention on source-side tokens. These results suggest that Transformer's tendency to process idioms as compositional expressions contributes to literal translations of idioms.

1 Introduction

An idiom is a group of words of which the figurative meaning differs from the literal reading, such as "kick the bucket," which means to die, instead of physically kicking a bucket. An idiom's figurative meaning is established by convention and is typically *non-compositional* – i.e. the meaning cannot be computed from the meanings of the idiom's parts. Idioms are challenging for the task of *neural machine translation* (NMT) (Barreiro et al., 2013; Isabelle et al., 2017; Constant et al., 2017; Avramidis et al., 2019). On the one hand, figures of speech are ubiquitous in natural language (Colson, 2019). On the other hand, idioms occur much less frequently than their parts, their meanings need to be memorised due to the non-compositionality,

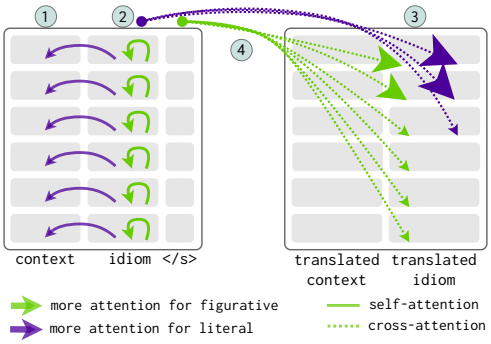


Figure 1: How do attention patterns of figurative PIEs that are paraphrased by the model compare to attention patterns of literal PIEs that are translated word for word? We find (1) decreased interaction between the PIE and its context, (2) increased attention within the PIE, (3) decreased cross-attention between the PIE and its paraphrase, (4) increased cross-attention from the paraphrase to </s>.

and they require disambiguation before translation. After all, not all *potentially idiomatic expressions* (PIEs) are figurative – e.g. consider "When I kicked the bucket, it fell over". Whether PIEs should receive a figurative or literal translation depends on the context. Yet, little is known about neural mechanisms enabling idiomatic translations and methods for improving them, other than data annotation (Zaninello and Birch, 2020). Related work studies how idioms are represented by Transformer-based language models (e.g. García et al., 2021a,b), but those models are not required to output a discrete representation of the idiom's meaning, which is a complicating factor for NMT models.

In this work, we analyse idiom processing for pre-trained NMT Transformer models (Vaswani et al., 2017) for seven European languages by comparing literal and figurative occurrences of PIEs. The comparison can help identify mechanics that underlie neural idiom processing to pave the way for methods that improve idiomatic translations. Large-scale analyses of idiom translations suffer

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3608 - 3626

from a lack of parallel corpora (Fadaee et al., 2018). We, therefore, use a monolingual corpus, heuristically label Transformer's translations, and verify the heuristic works as intended through human evaluation, as described in §3. To understand how idioms are represented in Transformer, we firstly apply interpretability techniques to measure the impact of PIEs on the encoder's self-attention and the cross-attention mechanisms (§4), as well as the encoder's hidden representations (§5). Afterwards, in §6, we intervene in the models while they process idiomatic expressions to show that one can change non-compositional translations into compositional ones.

The results indicate that Transformer typically translates idioms in a too compositional manner, providing a word-for-word translation. Analyses of attention patterns – summarised in Figure 1 – and hidden representations point to the encoder as the mechanism grouping components of figurative PIEs. Increased attention within the PIE is accompanied by reduced attention to context. When translating figurative PIEs, the decoder relies less on the encoder's output than for literal PIEs. These patterns are stronger for figurative PIEs that the model paraphrases than for sentences that receive an overly compositional translation and hold across the seven European languages. Considering that a recent trend in NLP is to encourage even more compositional processing in NMT (Raunak et al., 2019; Chaabouni et al., 2021; Li et al., 2021, i.a.), we recommend caution. It may be beneficial to evaluate the effect of compositionality-favouring techniques on non-compositional phenomena like idioms to ensure their effect is not detrimental.

2 Related Work

This section summarises work discussing human idiom comprehension, interpretability studies for NMT, and literature about figurative language processing in Transformer.

Idiom comprehension Historically, idioms were considered non-compositional units (Swinney and Cutler, 1979). Two main views (*literal first* and *direct access*) existed for how humans interpreted them. The former suggests humans attempt a compositional interpretation before considering the figurative interpretation in case of a contextual discrepancy (Bobrow and Bell, 1973; Grice, 1975, 1989). The latter view suggests one can immediately retrieve the non-compositional meaning (Gibbs Jr et al., 1994). The more recent *hybrid view* posits that idioms are simultaneously processed as a whole – primed by a *superlemma* (Kuiper et al., 2007) – and word for word (Caillies and Butcher, 2007). The processing speed and retrieval of the figurative meaning depend on the idiom's semantic properties and the context (Cain et al., 2009; Vulchanova et al., 2019). Examples of semantic properties are the conventionality and decomposability of idioms (Nunberg et al., 1994). We do not expect processes in Transformer to resemble idiom processing in humans. Nonetheless, this work helps us determine our focus of study on the role of the surrounding context and the extent to which idioms' parts are processed as a whole.

Translating PIEs that are used figuratively is not always straightforward. Baker et al. (1992) discuss strategies for human translators: (i) Using an idiom from the target language of similar meaning and form, (ii) using an idiom from the target language with a similar meaning and a different form, (iii) copying the idiom to the translation, (iv) paraphrasing the idiom or (v) omitting it. In the absence of idioms with similar meanings across languages, (iv) is the most common strategy. Our main focus is on literal translations (**word-for-word** translations), and **paraphrases**.

Interpreting Transformer Analyses of Transformer for NMT studied the encoder's hidden representations and self-attention mechanism (e.g. Raganato and Tiedemann, 2018; Tang et al., 2019b; Voita et al., 2019), the cross-attention (e.g. Tang et al., 2019a) and the decoder (e.g. Yang et al., 2020). The encoder is particularly important for the contextualisation of tokens from the source sentence; it acts as a feature extractor (Tang et al., 2019b). The encoder's bottom three layers better represent low-level syntactic features, whereas the top three layers better capture semantic features (Raganato and Tiedemann, 2018). As a result, one would expect the representations in higher layers to be more representative of idiomaticity.

Idioms are a specific kind of ambiguity, and whether a word is ambiguous can accurately be predicted from the encoder's hidden representations, as shown by Tang et al. (2019a) for ambiguous nouns. Transformer's cross-attention is not crucial for disambiguating word senses (Tang et al., 2018), but the encoder's self-attention does reflect ambiguity through more distributed attention for ambiguous nouns (Tang et al., 2019a).

Tropes in Transformer Various studies examine the Transformer-based language model BERT's (Devlin et al., 2019) ability to capture tropes like metonyms (Pedinotti and Lenci, 2020), idioms (Kurfalı and Östling, 2020), and multiple types of figurative language (Shwartz and Dagan, 2019). Kurfalı and Östling (2020) detect idioms based on the dissimilarity of BERT's representations of a PIE and its context, assuming that contextual discrepancies indicate figurative usage. Pedinotti and Lenci (2020) measure whether BERT detects meaning shift for metonymic expressions but find cloze probabilities more indicative than vector similarities. Shwartz and Dagan (2019) find that BERT is better at detecting figurative meaning shift than at predicting implicit meaning – e.g. predicting that "a hot argument" does not involve temperature.

The most recent work studies properties of hidden representations of noun-noun compounds (NCs) and verb-noun compounds (VCs): García et al. (2021b) examine (contextualised) word embeddings, including BERT, to compare figurative and literal NC *types*. They investigate the similarities between (1) NCs and their synonyms, (2) NCs and their components, (3) in-context and out-of-context representations, and (4) the impact of replacing one component in the NC. Surprisingly, idiomatic NCs are quite similar to their components and are less similar to their synonym compared to literal NCs. Moreover, the context of the NC hardly contributes to how indicative its representation is of idiomaticity, which was also shown by García et al. (2021a), who measured the correlation between *token*-level idiomaticity scores and NCs' similarity in- and out-of-context.

In search of the *idiomatic key* of VCs (the part of the input that cues idiomatic usage), Nedumpozhimana and Kelleher (2021) train a probing classifier to distinguish literal usage from figurative usage. They then compare the impact of masking the PIE to masking the context on the classifier's performance and conclude that the idiomatic key mainly lies within the PIE itself, although there is some information coming from the surrounding context.

3 Method

We use Transformer models (Vaswani et al., 2017) with English as the source language and one of seven languages as the target language (Dutch, German, Swedish, Danish, French, Italian, Spanish).[1] Transformer contains encoder and decoder networks with six self-attention layers each and eight heads per attention mechanism. The models are pre-trained by Tiedemann and Thottingal (2020) with the Marian-MT framework (Junczys-Dowmunt et al., 2018) on a collection of corpora (OPUS) (Tiedemann and Thottingal, 2020).[2] We extract hidden states and attention patterns for sentences with PIEs. The analyses presented are detailed for Dutch, after which we explain how the results for the other languages compare to Dutch.[3]

Parallel PIE corpora are rare, exist for a handful of languages only, and are limited in size (Fadaee et al., 2018). Rather than rely on a small parallel corpus, we use the largest corpus of English PIEs to date and annotate the translations heuristically. This section provides corpus statistics and discusses the heuristic annotation method.

MAGPIE corpus The MAGPIE corpus presented by Haagsma et al. (2020) contains 1756 English idioms from the Oxford Dictionary of English with 57k occurrences. MAGPIE contains identical PIE matches and morphological and syntactic variants, through the inclusion of common modifications of PIEs, such as passivisation ("the beans were spilled") and word insertions ("spill all the beans").[4] We use 37k samples annotated as fully **figurative** or **literal**, for 1482 idioms that contain nouns, numerals or adjectives that are colours (which we refer to as **keywords**). Because idioms show syntactic and morphological variability, we focus mostly on the nouns. Verbs and their translation are harder to identify due to the variability. Moreover, idiom indexes are also typically organised based on the nominal constituents, instead of the verbs (Piirainen, 2013). Only the PIE and its sentential context are presented to the model. We distinguish between PIEs and their context using the corpus's word-level annotations.

Heuristic annotation method The MAGPIE sentences are translated by the models with beam search and a beam size of five. The translations are labelled heuristically. In the presence of a literal translation of at least one of the idiom's keywords,

[1] Our figures refer to these languages using their ISO 639-1 codes, that are `nl`, `de`, `sv`, `da`, `fr`, `it` and `es`, respectively.

[2] The models are available via the `transformers` library (Wolf et al., 2020).

[3] The data and code are available via the `mt_idioms` github repository.

[4] Available via the `MAGPIE` github repository.

Category	nl	de	sv	da	fr	it	es
Figurative, paraphrase	20	20	24	18	19	20	24
Figurative, word for word	80	80	76	82	81	80	76
Literal, paraphrase	5	6	8	5	7	9	7
Literal, word for word	95	94	92	95	93	91	93

Table 1: Distribution of the heuristically assigned labels for translations of MAGPIE sentences in percentages, expressed within category (figurative / literal).

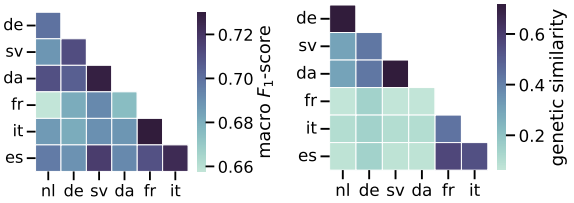


Figure 2: The macro-averaged F_1-score of translation labels (paraphrase vs word for word) for figurative PIEs and languages' genetic similarity visualised (Pearson's r=0.61, $p < 0.005$).

the entire translation is labelled as a **word-for-word** translation, where the literal translations of keywords are extracted from the model and Google translate. When a literally translated keyword is not present, it is considered a **paraphrase**.[5] Shao et al. (2018) previously analysed NMT translations of 50 Chinese idioms using a similar method and manually curated lists of literal translations of idioms' words to detect literal translation errors. Dankers et al. (2022) use a similar method for 20 English idioms, to track when a word-for-word translation changes into a paraphrased one during training for an English-Dutch (`En-Nl`) NMT model.

Table 1 summarises the distribution of these categories for all languages, for the subsets of figurative and literal examples from MAGPIE. Generally, paraphrased translations of figurative PIEs are more appropriate than word-for-word translations, whereas literal PIEs can be translated word for word (Baker et al., 1992). The vast majority of literal PIEs indeed result in word-for-word translations. The subset of figurative samples results in more paraphrases, but $\geq$ 76% is still a word-for-word translation, dependent on the language. Although the statistics are similar across languages, there are differences in which examples are paraphrased. Figure 2 illustrates the agreement

[5]The annotation does not evaluate whether paraphrases are correct, which requires expert idiom knowledge in both languages. A paraphrase being provided is a first step to adequately translating idioms and, at present, the only way to detect how the model approaches the task for large datasets.

Category	#	nl	de	sv	da	fr	it	es
Fig., paraphrase	116	88	84	75	81	78	78	87
Fig., word for word	103	95	92	95	74	96	97	82
Lit., paraphrase	28	54	71	43	82	43	32	50
Lit., word for word	103	98	89	97	89	98	100	94

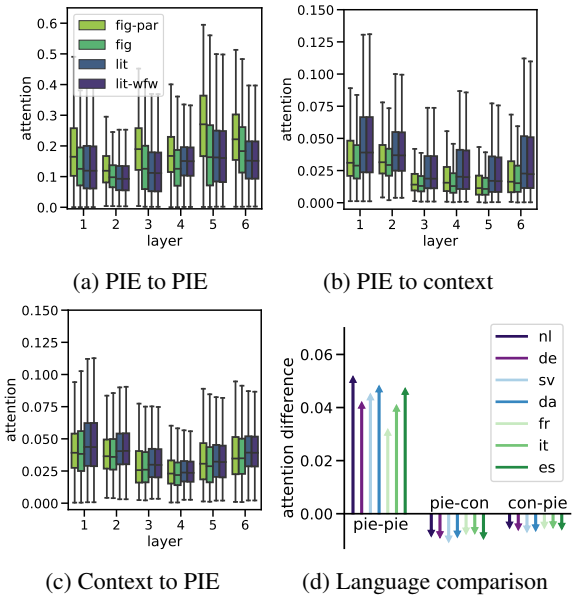
Table 2: Survey statistics: the number of sentence pairs used (#), and the percentage of labels for which the annotator and the algorithm agreed per language.

by computing the F_1-score when using the predictions for figurative instances of one language as the target, and comparing them to predictions from another language. The agreement positively correlates with genetic similarity as computed using the Uriel database (Littell et al., 2017).

To assess the quality of the heuristic method, one (near) native speaker per target language annotated 350 samples, where they were instructed to focus on one PIE keyword in the English sentence. Annotators were asked whether (1) the English word was present in the translation (initially referred to as "copy"), (2) whether there was a literal translation for the word, or (3) whether neither of those options were suited, referred to as the "paraphrase".[6] Due to the presence of cognates in the "copy" category, that category was merged with the "word for word" category after the annotation. Table 2 summarises the accuracies obtained. Of particular interest are samples that are figurative and paraphrased, since they represent the translations that are treated non-compositionally by the model, as well as instances that are literal and translated word for word, since they represent the compositional translations for non-idiomatic PIE occurrences. These categories have annotation accuracies of $\geq$ 75% and $\geq$ 89%, respectively. During preliminary analyses, an annotation study was conducted for Dutch by annotators from the crowd-sourcing platform Prolific. The annotators and the heuristic method agreed in 83% of the annotated examples, and for 77% of the samples an average of 4 annotators agreed on the label unanimously (see Appendix A for more details).

Sentences containing idioms typically yield lower BLEU scores (Fadaee et al., 2018). MAGPIE is a monolingual corpus and does not allow us to compute BLEU scores, but we refer the reader to Appendix G for an exploratory investigation for MAGPIE's idioms using the `En-Nl` training corpus.

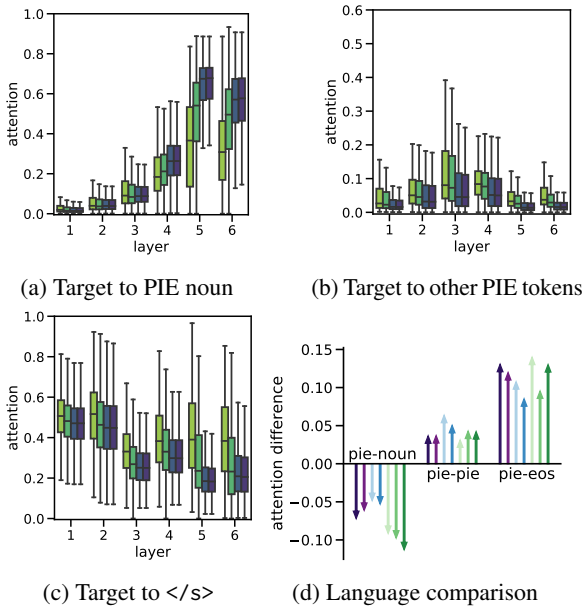
[6]Annotators were not involved in the research. Except for Swedish, annotators were native in the target language. For ethical considerations and more details, see Appendix A.



(a) PIE to PIE (b) PIE to context

(c) Context to PIE (d) Language comparison

Figure 3: Weight distributions of the encoder's self-attention (a-c), and the mean difference of *fig-par* and *lit-wfw* for all languages (d). Boxes represent quartiles; whiskers show the distribution, excluding outliers.



(a) Target to PIE noun (b) Target to other PIE tokens

(c) Target to </s> (d) Language comparison

Figure 4: The cross-attention for target-side tokens aligned to PIE nouns (a-c), and the mean difference between *fig-par* and *lit-wfw* for all languages (d).

4 Attention

We now turn to comparing how literal and figurative PIEs are processed by Transformer. Whether a PIE is figurative depends on the context – e.g. compare "in culinary school, I felt *at sea*" to "the sailors were *at sea*". Within Transformer, contextualisation of input tokens is achieved through the attention mechanisms, which is why they are expected to combine the representations of the idioms' tokens and embed the idiom in its context. This section discusses the impact of PIEs on the encoder's self-attention and the encoder-decoder cross-attention. To assert that the conclusions drawn in this section are not simply explained by shallow statistics of the data used, we recompute the results in Appendix C for (1) a data subset excluding variations of PIEs' standard surface forms, (2) a data subset that includes PIEs that appear in both figurative and literal contexts, (3) a data subset that controls for the number of tokens within a PIE. Qualitatively, these results lead to the same findings.

Attention within the PIE For the `En-Nl` Transformer, Figure 3a visualises the distribution of attention weights in the encoder's self-attention mechanism for incoming weights to one noun contained in the PIE from the remaining PIE tokens. Throughout the figures in the paper, we refer to the subset of sentences that have a figurative PIE and a paraphrased translation as '*fig-par*'. The subset of sentences with a literal PIE and a word-for-word translation are indicated by '*lit-wfw*'. We compare those two subsets, as well as all instances of figurative PIEs ('*fig*') to all instances of literal PIEs ('*lit*') using the labels from the MAGPIE dataset. Overall, there is increased attention in figurative occurrences of PIEs compared to literal instances. This difference is amplified for the subset of figurative PIEs yielding paraphrased translations. This pattern is consistent for all languages, as is displayed in Figure 3d that presents the difference between the mean attention weights of the figurative, paraphrased instances, and the mean weights of the literal instances translated word for word.[7] In other words, figurative PIEs are grouped more strongly than their literal counterparts.

Attention between PIEs and context To examine the interaction between a PIE and its context, we obtain the attention weights from tokens within the PIE to nouns in the surrounding context of size 10 (Figure 3b).[8] Similarly, the attention from the surrounding context to PIE nouns is measured (Figure 3c). There is reduced attention from PIEs to context for figurative instances, which mirrors the effect observed in Figure 3a: increased attention

[7] Appendix D details results per language per layer.

[8] Throughout the paper, a context size of 10 to the left and 10 to the right or smaller is used, as sentence length permits.

within the PIE is accompanied by reduced attention to the context. This pattern is consistent across languages (Figure 3d). From the context to the PIE, the average weight is slightly higher for literal PIEs, but the effect size is small, indicating only a minor impact of figurativeness on the context's attention weights. This will be further investigated in §5.

Cross-attention To analyse the encoder-decoder interaction, we decode translations with beam size five, and extract the cross-attention weights for those translations. Afterwards, alignments are computed for the models' predictions by, together with 1M sentences from the OPUS corpus per target language, aligning them using the `eflomal` toolkit (Östling et al., 2016). The alignment is used to measure attention from a token aligned to a PIE's noun to that noun on the source side.[9]

Figure 4a presents the attention distribution for the weights that go from the noun's translation to that PIE noun on the source side, for the `En-Nl` model. There is a stark difference between figurative and literal PIEs, through reduced attention on the source-side noun for figurative PIEs. This difference is particularly strong for the figurative sentences that are paraphrased during the translation: when paraphrasing the model appears to rely less on the source-side noun than when translating word for word. Where does the attention flow, instead? To some extent, to the remaining PIE tokens (Figure 4b). A more pronounced pattern of increased attention on the `</s>` token is shown in Figure 4c. Similar behaviour has been observed by Clark et al. (2019) for BERT's `[SEP]` token, who suggest that this indicates a *no-operation*. In Transformer's cross-attention mechanism, this would mean that the decoder collects little information from the source side. Figure 4d compares the mean attention weights of the seven languages for the figurative inputs that are paraphrased to the literal samples that are translated word for word, confirming that these patterns are not specific to `En-Nl` translation.

Collectively, the results provide the observations depicted in Figure 1. When paraphrasing a figurative PIE, the model groups idioms' parts more strongly than it would otherwise – i.e. it captures the PIE more as one unit. A lack of grouping all figurative PIEs could be a cause of too compositional translations. Increased attention within the PIE is accompanied by reduced interaction with context, indicating that the PIE is translated in a stand-alone manner, contrary to what is expected, namely that contextualisation can resolve the figurative versus literal ambiguity. There is less cross-attention on the source-side PIE and more attention on the `</s>` token when the model emits the translation of figurative (paraphrased) PIEs. This suggests that even though the encoder cues figurative usage, the decoder retrieves a PIE's paraphrase and generates its translation more as a language model would.

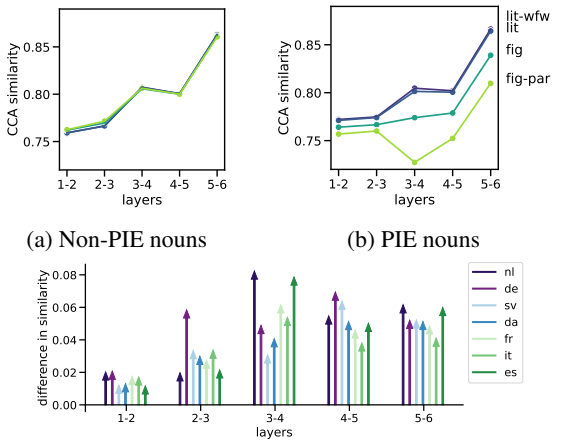
5 Hidden representations

Within Transformer, the encoder's upper layers have previously been found to encode semantic information (e.g. Raganato and Tiedemann, 2018). PIEs' hidden states are expected to transform over layers due to contextualisation, and become increasingly more indicative of figurativeness. This section focuses on the impact of PIEs on the hidden states of Transformer's encoder. We firstly discuss how much these hidden states change between layers. Secondly, we measure the influence of a token by masking it out in the attention and analysing the degree of change in the hidden representations of its neighbouring tokens. This analysis is performed to consolidate findings from §4, since the extent to which attention can explain model behaviour is a topic of debate (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

5.1 PIE changes over layers

To compare representations from different layers, we apply *canonical correlation analysis* (CCA) (Hotelling, 1936), using an implementation from Raghu et al. (2017). Assume matrices $A \in \mathcal{R}^{d_A \times N}$ and $B \in \mathcal{R}^{d_B \times N}$, that are representations for N data points, drawn from two different sources with dimensionalities d_A and d_B – e.g. different layers of one network. CCA linearly transforms these subspaces $A' = WA$, $B' = VB$ such as to maximise the correlations $\{\rho_1, \ldots, \rho_{\min(d_A, d_B)}\}$ of the transformed subspaces. We perform CCA using >60k random token vectors for a previously unused subset of the MAGPIE corpus – the subset of sentences that did not contain nouns in the PIEs –

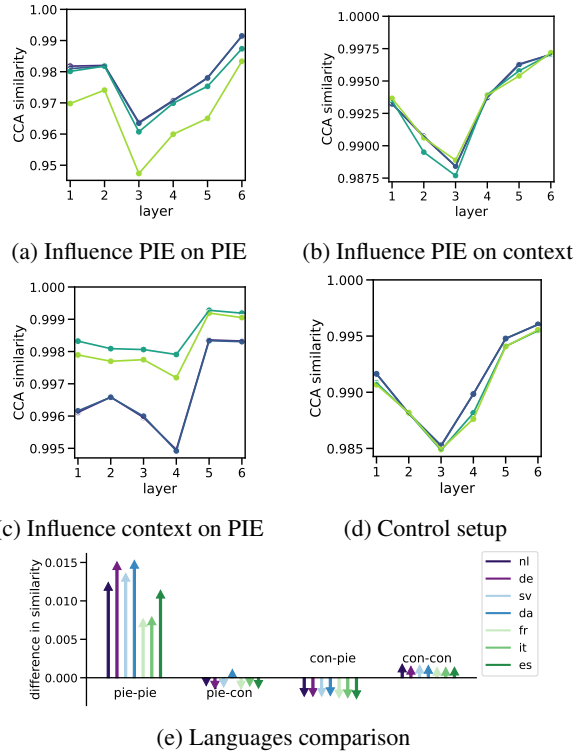
[9] Automated alignments may be less accurate for paraphrases, and, therefore, we inspect the *fig-par* alignments: for all languages $\leq$ 34% of those sentences has no aligned word for the PIE noun. Those sentences are excluded. We manually inspect the most frequently aligned words for Dutch, that cover 48% of the *fig-par* subcategory in Ap. B, and are all accurate.



(a) Non-PIE nouns (b) PIE nouns

(c) Languages comparison for PIE nouns

Figure 5: CCA similarity for layer l and layer $l+1$, for PIE and non-PIE nouns. The languages comparison displays the difference in similarity between *lit-wfw* and *fig-par*.



(a) Influence PIE on PIE (b) Influence PIE on context

(c) Influence context on PIE (d) Control setup

(e) Languages comparison

Figure 6: Impact of masking a PIE noun in the attention on (a) other PIE tokens, (b) other context tokens. Impact of masking a non-PIE noun on (c) PIE tokens and (d) other non-PIE tokens. (e) shows the difference in similarity between *lit-wfw* and *fig-par*.

to compute the CCA projection matrices W and V. W and V are then used to project new data points before measuring the data points' correlation. The CCA similarity reported in the graphs is the average correlation of projected data points. We do not perform CCA separately per data subset due to the small subset sizes and the impact of vocabulary sizes on CCA correlations for small datasets (see Appendix E).[10]

We compute the CCA similarity for hidden states from adjacent layers for PIE and non-PIE nouns. Figurative PIEs in layer l are typically less similar to their representation in layer $l-1$ compared to literal instances (shown in Figures 5b and 5c). The results for non-PIE nouns (Figure 5a for the `En-Nl` Transformer) do not differ across data subsets, suggesting that changes observed for figurative PIEs are indeed due to figurativeness.

5.2 Intercepting in attention

We now compute similarities of representations for the model in two setups: with and without one token masked in the attention mechanism, as suggested by Voita et al. (2019). Masking a token means that other tokens are forbidden to attend to the chosen one. This can reveal whether the attention patterns discussed in §4 are indicative of the influence tokens have on each other's hidden representations. The first representation is the hidden representation from layer l for a token encoded as usual. The second one is the hidden representation of layer l when applying the first $l-1$ layers as usual and masking one token in the lth layer. CCA is again performed on separate data, where a non-PIE noun is masked, to provide the projection matrices applied before computing similarities in the remainder of this subsection.

Masking a PIE token To estimate the influence of PIE nouns, we first compute the CCA similarity between two representations of tokens from the PIE's context while masking one PIE noun in the attention for one of those representations. Similarly, we measure the influence on other tokens within the PIE when masking one PIE noun. Within the PIE, the impact is the largest for figurative instances (see Figure 6a for `En-Nl` and 6e for averages over layers for all languages). This is in line with the attention pattern observed. However, whether the impact is the largest on context tokens from figurative or literal instances is dependent on the layer

[10] Extensions of CCA have been proposed that limit the number of CCA directions over which the correlation is computed, to only include directions that explain a large portion of the variance (Raghu et al., 2017; Morcos et al., 2018). We do not remove directions such as to avoid removing smaller variance components that could still cue figurativeness (the focus of our work).

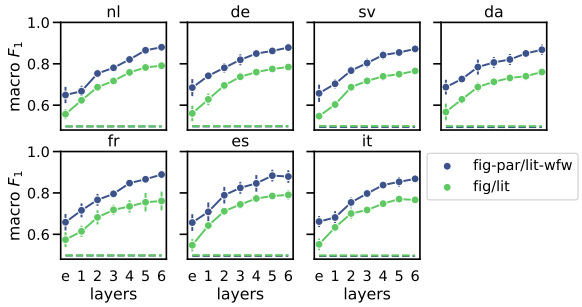
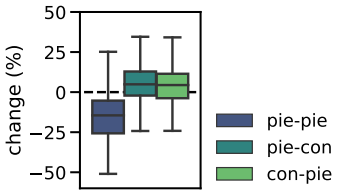


Figure 7: Macro F_1-score for probes predicting PIEs' labels. Error bars show standard deviations over folds.

	Fig.	Freq.
nl	36/75	34/75
de	33/68	33/69
sv	27/77	27/77
da	32/77	27/78
fr	37/77	30/76
it	39/76	34/77
es	40/78	30/78

(a) Success rate / BLEU



(b) Change in attention (nl)

Table 3: Impact of amnesic probing as (a) the success rate per PIE type (%), and the BLEU score of translations that changed from a paraphrase to a word-for-word translation, and (b) the changes in attention.

(Figure 6b), suggesting that the slight difference in attention from the context to the PIE observed in §4 need not represent a difference in impact between figurative and literal PIEs.

Masking a context token Lastly, we measure the influence of masking a noun in the context of the PIE on PIE tokens and non-PIE tokens. Within the PIE, as shown in Figures 6c and 6e, figurative instances are less affected by the masked context noun compared to literal occurrences of PIEs. Again, this mirrors the patterns observed for attention where there was less attention on the context for figurative PIEs. When masking a non-PIE noun and measuring the impact on non-PIE tokens, one would hardly expect any differences between data subsets, as is confirmed in Figures 6d and 6e.

In summary, these analyses confirm most of the trends noted for attention patterns. Intercepting in the attention through masking indicated that for PIE tokens, there is less interaction with the context. However, this does not necessarily mean that the context interacts less with figurative PIEs compared to literal PIEs, even if there was a slight difference in attention (see §4). The CCA analyses furthermore showed that figurative PIEs are distinct from typical tokens in how they change over layers.

6 (Amnesic) probing for figurativeness

The previous analyses compared the hidden states for figurative and literal PIEs, but do not use these labels, otherwise. We now train logistic regression *probing classifiers* (Conneau et al., 2018) to predict the label from hidden representations. The probes' inputs are the hidden states of PIE tokens, and the F_1-scores are averaged over five folds. All samples from one PIE are in the same fold, such that the classifier is evaluated on PIEs that were absent from its training data. The results (Figure 7) indicate figurativeness can be predicted from these encodings, with performance increasing until the top layer for all languages. F_1-scores for the embeddings already exceed a random baseline, indicating some idioms are recognisable independent of context.

Finally, we use probing classifiers to change models' PIE translations through *amnesic probing* (Elazar et al., 2021): removing features from hidden states with *iterative null-space projection* (INLP) (Ravfogel et al., 2020) and measuring the influence of these interventions. INLP trains k classifiers to predict a property from vectors. After training probe i, parametrised by W_i, the vectors are projected onto the nullspace of W_i. The projection matrix of the intersection of all k null spaces can then remove features found by these classifiers. Using INLP, we train 50 classifiers to distinguish figurative PIEs that will be paraphrased from those to be translated word for word. Afterwards, we run the previously paraphrased PIE occurrences through the model while removing information from the PIE's hidden states using INLP – i.e. information that could be captured by linear classifiers, which need not be the only features relevant to idiomatic translations. Per idiom, we record the percentage of translations that are no longer paraphrased. We report the scores for idioms from four folds and BLEU scores comparing translations that changed label before and after INLP. A fifth fold is used for parameter estimation (Appendix F).

Table 3 presents the results. When intervening in the hidden states for all layers $l \in \{0, 1, 2, 3, 4\}$, the average success rate per PIE ranges from 27% (for Swedish) to 40% (for Spanish). The interventions yield reduced attention within the PIE and increased interaction with the context (see Table 3b for Dutch). Table 3 also provides results for a baseline probe predicting whether the half-harmonic mean of the zipf-frequency of PIE tokens is below or above average. This probe is successful too,

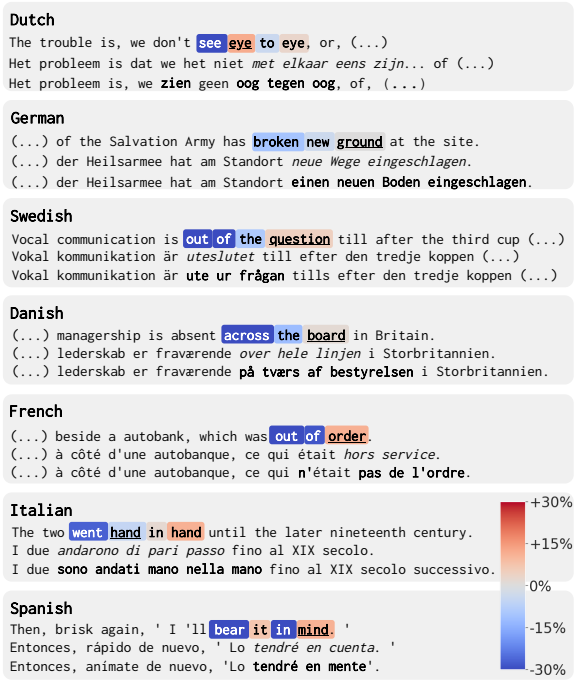


Figure 8: Source sentences and translations before and after INLP. PIEs and word-for-word translations are in bold font; paraphrases in italics. Colours indicate attention changes with respect to the underlined nouns.

emphasising how brittle idiomatic translations are: when removing information from the hidden states, the model reverts to compositional translations.

Figure 8 provides example translations before and after the application of INLP, while indicating how the attention on the underlined noun changes. Generally, the attention on that noun reduces for tokens other than itself.

In summary, when applying INLP to hidden states, the attention patterns resemble patterns for literal tokens more, confirming a causal connection between the model paraphrasing figurative PIEs and the attention. However, amnesic probing cannot change the paraphrases for all idioms; thus, figurativeness is not merely linearly encoded in the hidden states. The probing accuracies differed across layers and suggested figurativeness is more easily detectable in higher layers, which is in line with the changes across layers observed in §5.

7 Conclusion

Idioms are challenging for NMT models that often generate overly compositional idiom translations. To understand why this occurs, we analysed idiom processing in Transformer, using an English idiom corpus and heuristically labelled translations in seven target languages. We compared hidden states and attention patterns for figurative and literal PIEs. In the encoder, figurative PIEs are grouped more strongly as one lexical unit than literal instances and interact less with their context. The effect is stronger for paraphrased translations, suggesting that capturing idioms as single units and translating them in a stand-alone manner aids idiom processing. This finding agrees with results from Zaninello and Birch (2020), who ascertain that encoding an idiom as one word improves translations. It also agrees with the INLP application causing more compositional translations whilst changing the attention. By relying less on the encoder's output, the decoder determines the meaning of figurative PIEs more independently than for literal ones. To improve idiomatic translations, future work could use these insights to make architectural changes to improve the grouping of idioms as single units by training specific attention heads to capture multiword expressions or by penalising overly compositional translations in the training objective.

Although we learnt about mechanics involved in idiomatic translations, the vast majority of translations was still word for word, indicating that non-compositional processing does not emerge well (enough) in Transformer. Paradoxically, a recent trend is to encourage *more* compositional processing in NMT (Chaabouni et al., 2021; Li et al., 2021; Raunak et al., 2019, i.a.). We recommend caution since this inductive bias may harm idiom translations. It may be beneficial to evaluate the effect of compositionality-favouring techniques on non-compositional phenomena to ensure their effect is not detrimental.

Acknowledgements

We are grateful to Rico Sennrich for providing feedback on an earlier version of the paper. Many thanks to Agostina Calabrese, Matthias Lindemann, Gautier Dagan, Irene Winther, Ronald Cardenas, Helena Fabricius-Vieira and Emelie van de Vreken for data annotation and assistance with queries about their native languages. VD is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences. IT acknowledges the support of the European Research Council (ERC StG BroadSem 678254) and the

Dutch National Science Foundation (NWO Vidi 639.022.518).

References

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454.

Mona Baker et al. 1992. *In Other Words*. Routledge.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Fernando Batista, et al. 2013. When multiwords go bad in machine translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology*, pages 26–33.

Samuel A Bobrow and Susan M Bell. 1973. On catching on to idiomatic expressions. *Memory & cognition*, 1(3):343–346.

Stéphanie Caillies and Kirsten Butcher. 2007. Processing of idiomatic expressions: Evidence for a new hybrid view. *Metaphor and Symbol*, 22(1):79–108.

Kate Cain, Andrea S Towse, and Rachael S Knight. 2009. The development of idiom comprehension: An investigation of semantic and contextual processing skills. *Journal of experimental child psychology*, 102(3):280–298.

Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov. 2021. Can transformers jump around right in natural language? Assessing performance transfer from scan. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Jean-Pierre Colson. 2019. Multi-word units in machine translation: why the tip of the iceberg remains problematic–and a tentative corpus-driven solution. *Computational and Corpus-based Phraseology*, page 145.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: a neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marcos García, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741.

Marcos García, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564.

Raymond W Gibbs Jr, Raymond W Gibbs, and Jr Gibbs. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

H. Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:41–58.

H. Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 279–287.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Koenraad Kuiper, Marie-Elaine Van Egmond, Gerard Kempen, and Simone Sprenger. 2007. Slipping on superlemmas: Multi-word lexical items in speech production. *The Mental Lexicon*, 2(3):313–357.

Murathan Kurfalı and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 85–94.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Ari S Morcos, Maithra Raghu, and Samy Bengio. 2018. Insights on representational similarity in neural networks with canonical correlation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5732–5741.

Vasudevan Nedumpozhimana and John Kelleher. 2021. Finding BERT's idiomatic key. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 57–62.

Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Robert Östling, Jörg Tiedemann, et al. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Paolo Pedinotti and Alessandro Lenci. 2020. Don't invite BERT to drink a bottle: Modeling the interpretation of metonymies using BERT and distributional representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6831–6837.

Elisabeth Piirainen. 2013. Widespread idioms in europe and beyond. toward a lexicon of common figurative units. *Neuphilologische Mitteilungen*, 13(4):489–.

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. *EMNLP 2018*, page 287.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6078–6087.

Vikas Raunak, Vaibhav Kumar, Florian Metze, and Jaimie Callan. 2019. On compositionality in neural machine translation. In *NeurIPS 2019 Context and Compositionality in Biological and Artificial Neural Systems Workshop*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. 2018. Evaluating machine translation performance on chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

David A Swinney and Anne Cutler. 1979. The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*, 18(5):523–534.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019a. Encoders help you disambiguate word senses in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1429–1435.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019b. Understanding neural machine translation by simplification: The case of encoder-free models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1186–1193.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4387–4397.

Mila Vulchanova, Evelyn Milburn, Valentin Vulchanov, and Giosuè Baggio. 2019. Boon or burden? The role of compositional meaning in figurative language processing and acquisition. *Journal of Logic, Language and Information*, 28(2):359–387.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yilin Yang, Longyue Wang, Shuming Shi, Prasad Tadepalli, Stefan Lee, and Zhaopeng Tu. 2020. On the sub-layer functionalities of transformer decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4799–4811.

Andrea Zaninello and Alexandra Birch. 2020. Multiword expression aware neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3816–3825.

Appendix A Survey details

A.1 Crowd-sourcing annotations for Dutch

In an early phase of the research, the quality of the heuristic annotation method was estimated through a survey conducted using the Qualtrics platform by annotators from Prolific. The heuristic annotation method labelled a translation as 'word for word' if the literal translation of a keyword was present, where the keyword was elicited from MarianMT, and from the translation tool DeepL. These annotators were native speakers of Dutch, and fluent in English. To guard the quality of the data collection, participants went through a pre-screening process that consisted of a shorter version of the survey with three practice questions and seven regular questions. Participants were selected for the full study if they correctly answered practice questions, used all three of the labels (paraphrase, word for word, copy), and did not choose 'copy' if the keyword was clearly absent from the translation. The main survey consisted of three parts: (1) An explanation of what an idiom is, of potential literal and figurative usage of PIEs, the meaning of the three labels, and the format to be used in the study. (2) One practice exercise where three potential translations of one sentence had to be connected to the correct label. (3) Lastly, 38 questions were filled out: 12 instances that were figurative and were paraphrased by the model, 4 literal instances paraphrased by the model, 8 literal instances that were translated word for word, 8 figurative instances that were translated word for word, 6 copies (3 figurative, 3 literal).

If the participant indicated that it was a word-by-word translation, the follow-up question would be asked, where the participant indicated the literal translation of the keyword. We repeated the instruction of what constitutes a word-by-word translation since participants would often select the (conventionalised) idiomatic translation in the pre-screening phase – e.g. 'handbereik' for 'fingertips', for which a literal translation would be 'vingertoppen'.

Table 4 summarises the survey outcomes. The annotators and the heuristic method agreed in 83% of the cases. For 77% of the samples, the annotations agreed on the label unanimously.

A.2 Collecting annotations for 7 languages

Later on, the analyses were applied to heuristically annotated data for all seven languages. The procedure to elicit the translations of keywords from MarianMT and an online translation tool were adapted to improve the recall of keywords for languages other than Dutch. Afterwards, postgraduate students from the local university were invited to annotate the data in exchange for payment, where one annotator annotated all 350 samples for a language. To reduce the cognitive load of the experiment, only sentences with ≤ 40 tokens were shown to the participants. The annotators were native in the target language and fluent in English, with the exception of the Swedish speaker, that was native in Norwegian and Finnish, and fluent in Swedish and English. The annotators participated in a similar pre-screening test with language-specific explanations and examples, and seven practice questions. If the annotators' answers differed from what was expected, the instructions were discussed with the annotator before they proceeded with the full survey, and they filled out the remainder of the survey without intermediate help or instructions. Table 5 shows an example question for Dutch.

MAGPIE	Predicted Translations								
	Paraphrase			*Word for word**			*Copy**		
	#	%	agr	#	%	agr	#	%	agr
Figurative	96	86	84	64	84	77	24	83	58
Literal	32	73	59	64	91	80	24	69	88

Table 4: Survey statistics: the number of sentence pairs used (#), the % of labels the algorithm and annotators agreed on, and inter-annotator agreement. Agreement means an average of 4 annotators agreed on the label unanimously. *Categories merged in the main paper.

Question
The following sentence contains "at your fingertips":
"Using the latest in audio visual technology, the wonders of these six fascinating 'worlds' are at your fingertips."

Now categorise the translation of the red word from above in this sentence:
"Met behulp van de nieuwste audio visuele technologie, zijn de wonderen van deze zes fascinerende werelden binnen handbereik."
○ paraphrase
○ word-by-word
○ copy

Follow-up question
If you did not select 'word for word', leave blank.
What is the translation of the red keyword in "at your fingertips" in the sentence below:
(. . . insert sentence. . .)
(. . . free text response box. . .)

Table 5: Format of the questions shown to participants via the Qualtrics platform.

A.3 Ethical considerations

The surveys referred to in §3 were both approved through to the university's research ethics process,

where an independent committee assessed the setup of the survey, the research's potential harmful impacts and the compensation for the participants. In collecting data annotations, participants were shown data from the MAGPIE corpus, available under the CC-BY-4.0 License. All other information shown to them was either collected from the computational model, or written up by the authors. Any identifiable information about the participants was stored separately from the participants' annotations, for the purposes of compensation. Participants were able to provide informed consent to data collection and anonymised data being used in academic publications. They were given the opportunity to withdraw at any time. Participants were compensated above the minimum hourly wage of the country in which they were a resident at the time of participating in the study.

Appendix B Aligning PIEs and paraphrases

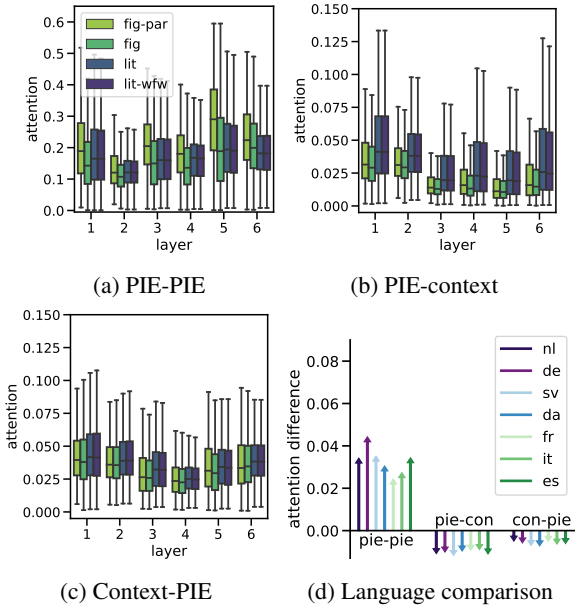
When automatically aligning sentences with PIEs to translations that are labelled as a paraphrase by the heuristic, how does the automated aligner (the `eflomal` toolkit of Östling et al., 2016) handle paraphrases? For many PIEs ($\leq$ 34% of the *fig-par* sentences for all languages), the paraphrases do not have a word in the translation aligned to the PIE keyword on the source side using `eflomal`. These examples are excluded. However, for a subset that appears more well-known, there are common paraphrases that the PIE keyword aligns to. We provide examples for Dutch in Table 6. The examples provided in the table together cover 48% of all aligned sentences used in the cross-attention analysis for the *fig-par* category, and all are reasonable alignments.

PIE	**Dutch paraphrase (literal backtranslation)**	**Aligned tokens**
across the board	over hele linie (*over the whole line*)	board → linie
behind the scenes	achter de schermen (*behind the screens*)	scenes → schermen
break new ground	nieuwe weg inslaan (*take a new road*)	ground → weg
by heart	uit het hoofd (*from the head*)	heart → hoofd
by the same token	op dezelfde manier (*in the same way*)	token → manier
come to mind	in me opkomen (*come up in me*)	mind → me
come of age	volwassen worden (*become an adult*)	age → volwassen
face to face	oog in oog (*eye in eye*)	face → oog
follow suit	het voorbeeld volgen van (*follow the example of*)	suit → voorbeeld
for good measure	in goede mate* (*in good measure*)	measure → mate
from scratch	vanaf nul (*from zero*)	scratch → nul
from the word go	vanaf het begin (*from the start*)	word → begin
get a move on	schiet op (*hurry*)	move → schiet
get the picture	een completer beeld krijgen (*get a more complete vision*)	picture → beeld
get to grips with	(aan)pakken (*take on*)	grips → pakken
give someone the creeps	kriebels krijgen (*getting tickles*)	creep → (krie)bel
in broad daylight	op klaarlichte dag (*on a luminous day*)	day(light) → dag
in full swing	in volle gang (*in full progress*)	sw(ing) → gang
in the flesh	in levende lijve (*in the living body*)	flesh → lij(ve)
in the long run	op de lange termijn (*on the long term*)	run → termijn
in the short run	op de korte termijn (*on the short term*)	run → termijn
keep a low profile	zich gedeisd houden (*to lay low*)	profile → (gede)is(d)
off the record	onofficieel (*unofficial*)	record → (onoffici)eel
on someone's mind	iets aan je hoofd hebben (*have something on your head*)	mind → hoofd
once in a while	af en toe (*on and off*)	while → toe
out of the blue	uit het niets (*out of nothing*)	blue → niets
out of the question	uit de boze (*from the bad*)	question → boze
set eyes on	zien / zag (*see / saw*)	eyes → zag
small print	in de kleine lettertjes (*in the little letters*)	print → (letter)tjes
take a back seat	op de achterbank (*on the back bench*)	seat → bank
take stock	de balans opmaken (*make up the balance*)	stock → balans
to all intents and purposes	in alle opzichten (*in all aspects*)	intent → opzichten
to boot	opstarten (*to start*)	boot → (op)starten
to the tune of	voor het bedrag van (*for the amount of*)	tune → bedrag
with a view to	met het oog op (*with the eye on*)	view → oog

Table 6: PIEs for which the word most commonly aligned to the keyword occurs $>$ 20 times. Together, these keywords determine 48% of all the alignments used to perform the cross-attention analysis for *fig-par* in the English-Dutch model. Subwords shown in brackets are due to the subtokens used in Marian-MT: `eflomal` aligns the parts outside of the brackets to one another.
*Example of a PIE for which the heuristic annotation missed out on a potential literal translation of 'measure'.

Appendix C Attention for data subsets

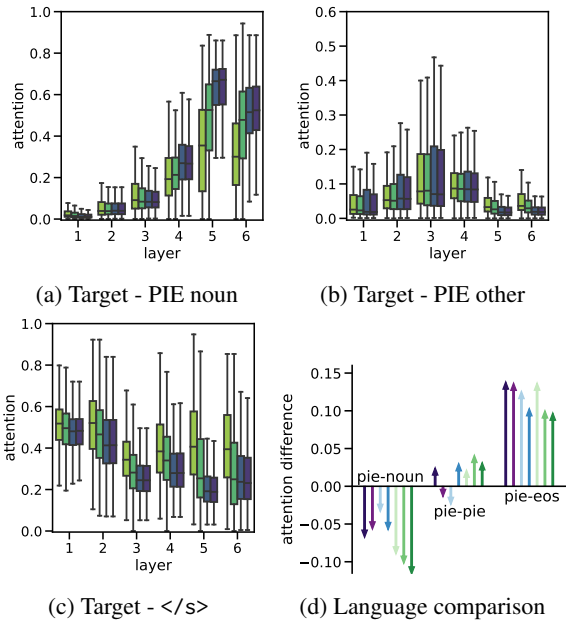
The attention weight distributions in the main paper included all data. To further investigate whether the differences in attention patterns observed are due to factors other than figurativeness, we recompute the attention patterns for three additional data subsets.



(a) PIE-PIE (b) PIE-context

(c) Context-PIE (d) Language comparison

Figure 9: Encoder self-attention distributions, illustrating attention within the PIE and the interaction between the PIE and its context, for the identical data subset.



(a) Target - PIE noun (b) Target - PIE other

(c) Target - </s> (d) Language comparison

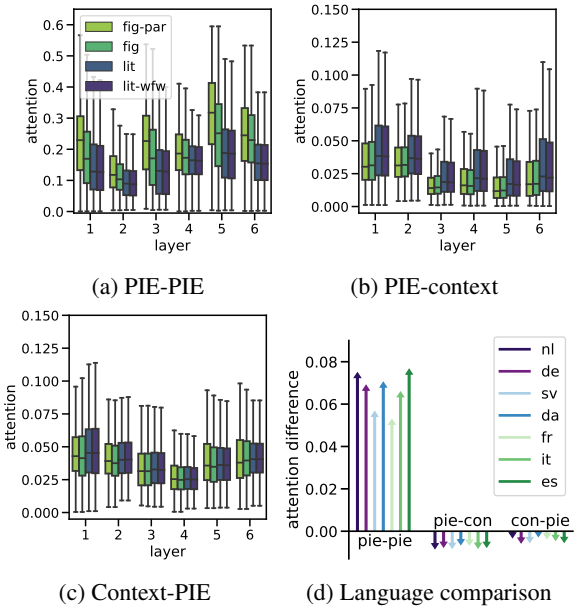
Figure 10: Cross-attention distributions from the translation of one PIE noun on the target side to that noun on the source side, for the identical data subset.

PIE identical matches We first use a subset that only includes samples for which MAGPIE reports an **identical match** between the PIE and the English sentence, that includes 17k samples. This subset excludes sentences with modifications to the typical surface form of a PIE, such as upper-cased tokens or insertions of a token into the PIE (e.g. "That gossip of a man spilled *all* of the beans.").

Figure 9 shows the three attention patterns previously discussed for the encoder's self-attention – i.e. attention from the PIE to the PIE, attention from the PIE to the context, and from the context to the PIE. Overall, the patterns resemble those discussed in the main text, apart from Figure 9a, where figurative instances do not display consistently higher attention weights compared to literal instances, although the *fig-par* subset does.

This procedure is repeated for the cross-attention distributions. Figure 10 depicts the three patterns discussed in the main paper – i.e. from the aligned target-side tokens to the PIE noun, to another PIE token, and to </s> – for this data subset, providing the same qualitative findings.

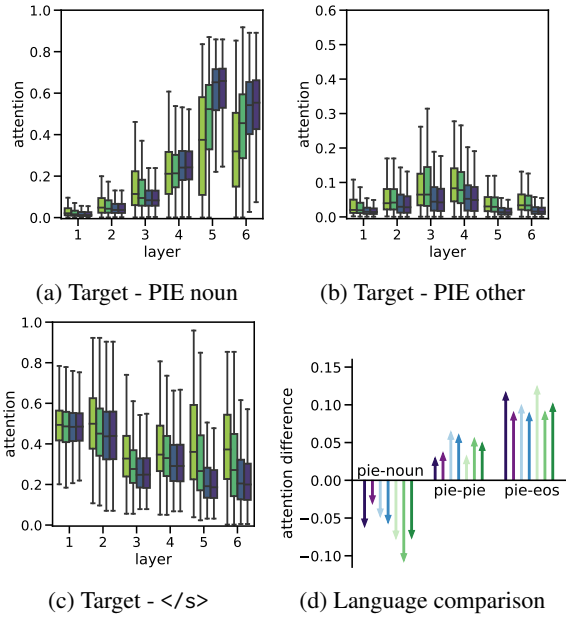
Intersection of PIEs The second subset (referred to as **intersection**) considered is one that only contains idioms that are among all of the subsets of figurative, literal, paraphrased and word for word instances, covering 11k examples from the dataset. The results for the encoder's self-attention



(a) PIE-PIE (b) PIE-context

(c) Context-PIE (d) Language comparison

Figure 11: Encoder self-attention distributions, showing attention within the PIE and the interaction between the PIE and its context, for the intersection data subset.



(a) Target - PIE noun (b) Target - PIE other

(c) Target - </s> (d) Language comparison

Figure 12: Cross-attention distributions from the translation of one PIE noun on the target side to that noun on the source side, for the intersection data subset.

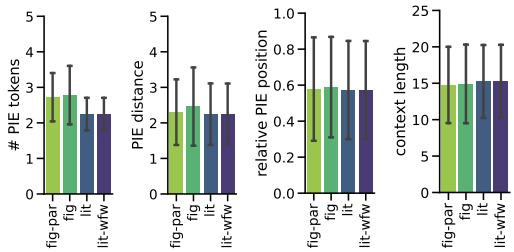
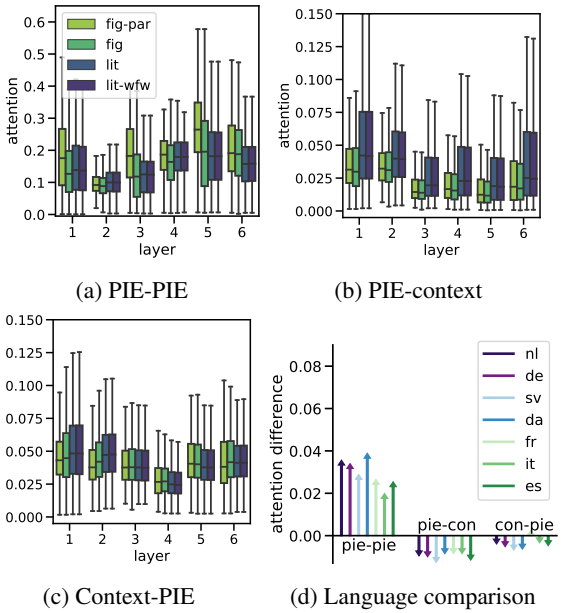


Figure 13: Length statistics for the four categories of PIEs (*fig-par*, *lit-wfw*, *fig*, *lit*). Error bars indicate standard deviations over sentences.



(a) PIE-PIE (b) PIE-context

(c) Context-PIE (d) Language comparison

Figure 14: Encoder self-attention distributions, showing attention within the PIE and the interaction between the PIE and its context, for the length controlled subset.

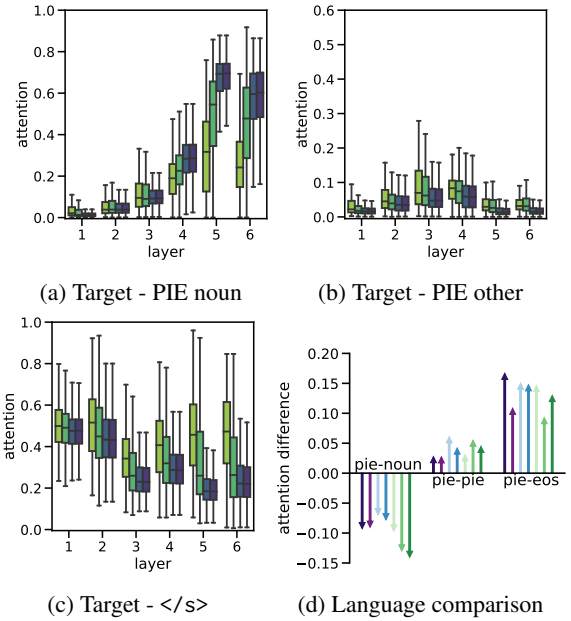
patterns are shown in Figure 11. Figure 12 summarises the results for the cross-attention mechanisms. These results lead to the same qualitative findings as mentioned in the main paper, and, in the encoder, the PIE to PIE attention patterns for figurative and literal PIEs are even more distinct.

Controlling PIE length To investigate the impact of the length of a PIE and the length of its context on the results, we now report additional measures over sentences, namely:

- the **average number of MarianMT tokens** labelled as being part of the PIE (in MAGPIE words like prepositions and determiners are not counted as part of the PIE, so the annotation can be discontinuous);
- the **distance between the first and the last token** of the PIE (two tokens right next to each other have a distance of 1);
- the **relative position** of the tokens that are annotated as belonging to the PIE, which impacts the potential context size, but could also impact how a PIE 'behaves';
- the average distance of the first position of the PIE's context tokens (PIE - 10) to the last position (PIE + 10) (**context length**).

Figure 13 summarises these statistics for the MAGPIE PIEs. The last two metrics are very stable across categories, with an average relative position of 0.57 for PIEs (0.58 for figurative, 0.56 for literal), and context lengths of 17.0 (17.0 for figurative, 17.1 for literal). The first two metrics indicate that figurative PIEs are a bit longer than literal PIEs (0.69 words), and that the distance between the first and the last word is slightly larger (0.46 positions).

To assert that these differences do not substantially impact our qualitative findings, we compute the attention analyses over a data subset that only uses sentences where there are three tokens annotated for the PIE, for which the start and end are three positions apart. This covers a subset of approximately 7k samples, with small variations between languages due to slightly different tokenisation of the English words. Figures 14 and 15 present the results for the encoder's self-attention and the encoder-decoder cross-attention analyses, respectively. Qualitatively, our findings for this subset do not differ from the previous findings.



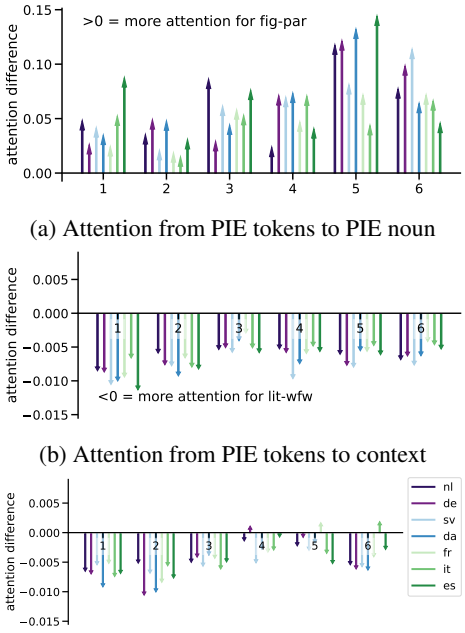
(a) Target - PIE noun (b) Target - PIE other

(c) Target - </s> (d) Language comparison

Figure 15: Cross-attention distributions from the translation of one PIE noun on the target side to that noun on the source side, for the length controlled subset.

Appendix D Results for 7 languages, per layer

Figures 16 and 17 present the results per layer, for the (cross-)attention graphs from §4. Figure 18 present the results per layer, for the CCA similarity graphs from §5.

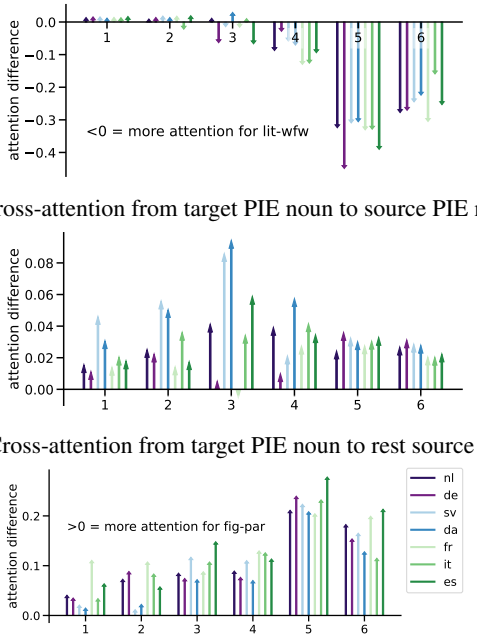


(a) Attention from PIE tokens to PIE noun

(b) Attention from PIE tokens to context

(c) Attention from context to PIE

Figure 16: The differences in attention between *fig-par* and *lit-wfw* visualised per layer, per language.

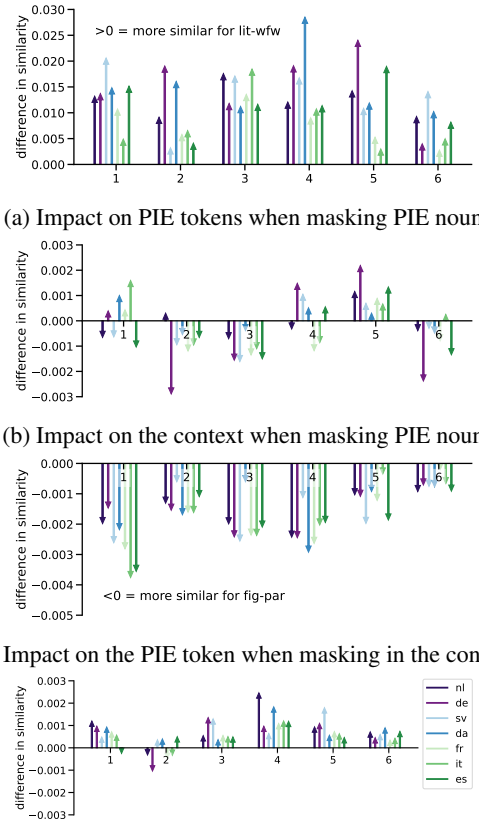


(a) Cross-attention from target PIE noun to source PIE noun

(b) Cross-attention from target PIE noun to rest source PIE

(c) Cross-attention from target PIE noun to </s>

Figure 17: The differences in cross-attention between *fig-par* and *lit-wfw* visualised per layer, per language.



(a) Impact on PIE tokens when masking PIE noun

(b) Impact on the context when masking PIE noun

(c) Impact on the PIE token when masking in the context

(d) Impact on the context token when masking in the context

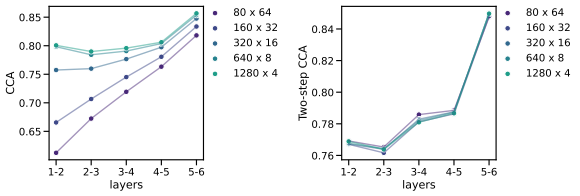
Figure 18: The differences in CCA similarity between *lit-wfw* and *fig-par* visualised per layer, per language. Here, "more similar" means that the impact of masking is smaller.

Appendix E Two-step CCA

CCA can be used to compare representations over different layers of the same network or different networks in a way that is invariant to affine transformations (Raghu et al., 2017). The CCA similarity expresses the extent to which two representations contain the same information while accounting for transformations in these two views of the data. Nonetheless, the similarity depends on the data used to perform CCA. Even with a dataset that is at least an order of magnitude larger than the number of dimensions in the hidden representations, the composition of the dataset impacts the outcome. Particularly relevant in the context of our work is the vocabulary size that impacts CCA computations.

We illustrate this by measuring how hidden representations change over layers, randomly sampling tokens and considering multiple dataset compositions, varying from 64 occurrences of 80 unique tokens, to 4 occurrences of 1280 unique tokens. Recomputing CCA per subset yields the similarities shown in Figure 19a. Although the overall pattern of lower similarity between lower layers and higher similarity between higher layers is present for all subsets, the absolute similarity measures differ between subsets. In Figure 19b, however, where the projection matrix is computed on a separate dataset, subsets show comparable similarities. The differences between the methods decrease as the number of hidden representations used to perform CCA grows.

Performing CCA separately per (relatively small) subset of the MAGPIE corpus could thus reflect vocabulary differences rather than systematic differences due to figurativeness. We merely want to apply CCA to account for differences between layers and differences with and without masking attention, and thus apply two-step CCA, computing projection matrices on a separate dataset.



(a) CCA per subset (b) Two-step CCA

Figure 19: Illustration of the impact of recomputing CCA with data subsets of differently composed vocabularies for a dataset size of 5k.

Appendix F Amnesic probing

Amnesic probing (Elazar et al., 2021) evaluates the behavioural influence of information recovered from hidden representations H by probes, by removing that information from the representation and measuring the change in behaviour on the main task. INLP, proposed by Ravfogel et al. (2020), is used to remove this information from the representations, by training k classifiers to predict a property from input vectors. After training probe i, parametrised by W_i, the vectors are projected onto the nullspace of W_i, using projection matrix $P_{N(W_i)}$, such that $W_i P_{N(W_i)} H = 0$. The projection matrix of the intersection of all k null spaces can then remove features found by the k classifiers.

Using INLP, we train 50 classifiers to detect figurative, paraphrased PIEs from figurative PIEs translated word for word from the hidden state. Afterwards, we apply the projection matrices while the model processes previously paraphrased translations. We separate the PIEs into five folds, using one for parameter estimation. For every fold $\frac{3}{5}$ is used to train INLP's probes, $\frac{1}{5}$ is used to measure whether the performance of the k probes decreases and $\frac{1}{5}$ is used to measure the changed percentage. Dependent on where one intervenes in the model, amnesic probing may be more or less successful, since not every layer encodes the linguistic property and higher layers could recover information removed from lower layers (Elazar et al., 2021). The parameter estimation performed measures the impact of different combinations of layers as the average success rate per PIE type (success means achieving a word-for-word translation). As shown in Figure 20, there is quite some variation among languages, but generally intervening in the lower layers of Transformer is the most successful, and including the sixth layer is quite detrimental. The results in the main body of the paper are computed

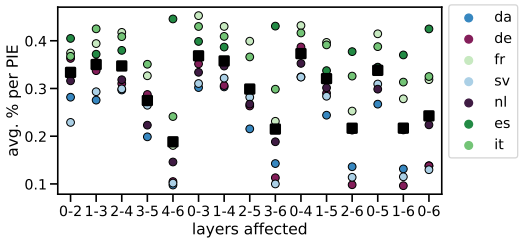


Figure 20: Impact of the selection of layers affected by INLP. Dots represented different languages, the squares indicate the mean %.

by intervening on the hidden states of PIE tokens in $l \in \{0, 1, 2, 3, 4\}$.

Appendix G Idioms in OPUS

To understand whether the model's translations reflect target translations from its training corpus, we extract up to 500 identical matches per idiom from OPUS for the `En-Nl` model. These target translations are labelled heuristically, resulting in 54% of paraphrased instances, which is substantially higher than the percentage of paraphrased instances in the model's translations. This may be the result of infrequent idioms contained in OPUS, for which the model fails to learn the correct implicit meaning, even though the corpus does provide paraphrases. Table 7 illustrates how the predicted translations' labels relate to the labels of target translations and provides BLEU scores per subset. Samples with a paraphrased target translation score substantially lower compared to those with a word-for-word or copied target translation, emphasising the negative impact of idioms on translation quality.

OPUS		**Predicted Translations**	
		Paraphrase	*Word for word*
(a) Translation type frequency (%)			
Paraphrase	54	49	51
Word for word	46	7	93
(b) BLEU scores			
Paraphrase		27.2	19.9
Word for word		25.6	38.2

Table 7: Distribution of translation labels for idiom occurrences in OPUS, along with their BLEU scores.