Memorisation Meets Compositionality in Natural Language Processing

Verna Dankers
University of Edinburgh
McGill University, Mila – Quebec AI Institute
vernadankers@gmail.com

Abstract

Memorisation in deep learning is undergoing a paradigm shift; it is increasingly recognised as a mechanism that can support, rather than hinder, generalisation. This is particularly relevant in NLP, where language combines compositional, generalisable structure with non-compositional expressions such as idioms, requiring memorisation from models and humans alike. My PhD work investigated memorisation in transformer models in generic terms, and through the lens of (non-)compositionality, from both data and model-internal perspectives. I analysed which training examples require memorisation, whether memorisation supports generalisation, and where memorisation occurs within model layers. I also studied how transformers process non-compositional idiom translations and how they balance compositional generalisation with non-compositional memorisation. Based on my findings, I stress that memorisation is an inherent part of learning *natural* language, can be beneficial, and is partially predictable. Yet it is not cleanly separable from generalisation, both at the level of data and of model parameters. Here, I summarise those findings and reflect on my PhD work.

1 Introduction

In deep learning, the perspective on memorisation of training examples is undergoing a paradigm shift. Previously linked to overfitting and poor generalisation, memorisation is now seen both as beneficial when it enhances deep neural networks' generalisation capabilities (e.g. Feldman, 2020; Feldman and Zhang, 2020; Zheng and Jiang, 2022) and as concerning when it involves examples that should not be memorised (e.g. Huang et al., 2022; Chang et al., 2023). This shift raises questions about how much models *can* even memorise, what they should or should not memorise, and how memorisation is implemented internally. Although these questions apply broadly, they are particularly relevant for language learning and NLP.

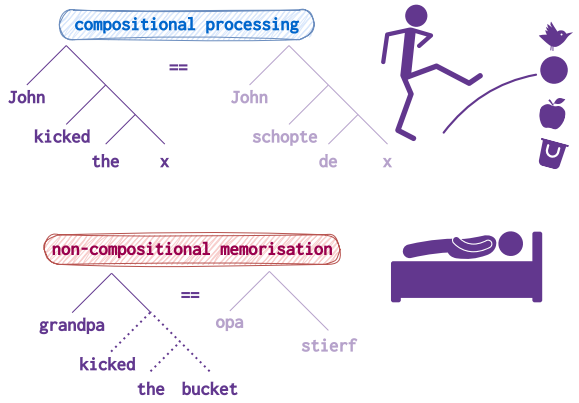


Figure 1: Illustration of the relation between compositionality and memorisation. "<person> kicked the <object>" is normally processed compositionally, yet "kicked the bucket" needs to be memorised as a single unit. This affects NLP tasks, such as translation.

Natural language itself requires both syntax-driven, generalisable meaning composition *and* memorisation, because it is simultaneously compositional and non-compositional – due to the prevalence of fixed formulaic expressions, among which proverbs, idioms and non-compositional noun compounds (Wray, 2002; Baggio, 2021). Many non-compositional expressions cannot be interpreted by composing the meanings of their parts, so they must be stored and retrieved holistically. Without such memorisation, humans and NLP models alike would default to literal readings (e.g. interpreting "grandpa kicked the bucket" literally rather than as "passed away" in Figure 1). The non-compositional side of language makes memorisation an essential complement to compositional processing in natural language understanding.

In my dissertation, I investigated memorisation in computational models of language, both as a

Home institute while conducting the PhD work.
Home institute at the time of submission.

Proceedings of The Big Picture v2: Crafting a Research Narrative, pages 144–159

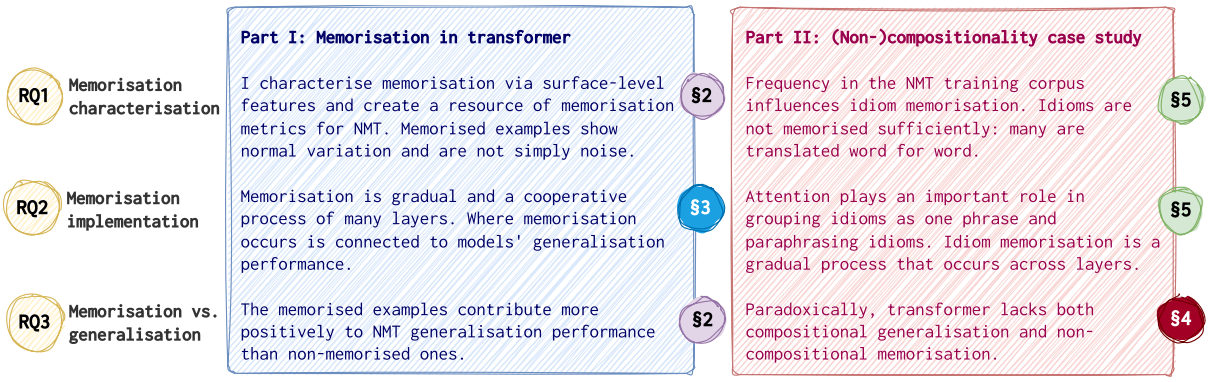


Figure 2: An overview of how the different sections address the three central research questions, both for memorisation in general (in blue) and for a (non-)compositionality case study (in red).

general phenomenon and through the lens of non-compositionality, examining the compositional vs non-compositional dichotomy as a memorisation–generalisation case study. In the process, both *neural machine translation* (NMT) and generic natural language understanding classification tasks were used. All analyses adopted the transformer architecture (Vaswani et al., 2017), albeit in different setups, varying training set sizes, model sizes and whether or not the model was pretrained. In this write-up presented to you here, I summarise my dissertation's work, highlighting findings and providing a retrospective.[1] This write-up is divided into two parts, each containing two sections with experiments and findings drawn from two previously published papers, together addressing the following research questions:

1. What characterises memorised examples? (§2 for generic memorisation patterns, §5 for idioms, specifically)
2. Which model-internal mechanisms enable memorisation? (§3 for memorised mislabelled examples, §5 for idioms)
3. To what extent are memorisation and generalisation at odds with one another? (§2 for how counterfactual memorisation benefits generalisation, §4 and §5 for the balance between compositional generalisation and non-compositional memorisation)

Lastly, §6 summarises the answers to these questions, which are also illustrated in Figure 2. Furthermore, Appendix A summarises lessons learnt during the PhD. Appendix B elaborates on PhD work that was not incorporated in the thesis.

[1]The dissertation extended the experiments conducted in the original papers. Findings that are the result of those added experiments are indicated in Sepia.

Part I
Memorisation in transformer

I will first focus on memorisation in generic terms: within a dataset, some examples require more memorisation than others. Which examples do models memorise, and where in these multi-layered networks can memorisation be localised?

2 A memorisation-generalisation continuum of data (Dankers et al., 2023)

When training neural networks, we aim for models to generalise rather than simply memorise training data. However, fitting natural language datasets inevitably requires memorising some of the data's idiosyncrasies (e.g. Feldman, 2020; Zheng and Jiang, 2022; Zhang et al., 2023). That memorisation is not always harmful, and can benefit generalisation had previously been established prior to the work discussed in this section, but primarily for artificial setups or classification tasks (Feldman and Zhang, 2020; Raunak et al., 2021; Zheng and Jiang, 2022).

Yet, is memorisation still beneficial in the *real-world*, noisy domain of NMT? Very little prior work has discussed memorisation in the context of NMT, with the exception of Raunak et al. (2021); Raunak and Menezes (2022). These works, however, focused on memorisation in a narrow sense, applied to a very small set of examples, and discussed it in relation to hallucinations. We, on the other hand, perform a very comprehensive analysis by constructing a multilingual resource of memorisation metrics, analysing which datapoint characteristics influence memorisation, and examining how memorisation relates to performance.

2.1 Experiments and findings

We treat memorisation as a graded phenomenon, quantified using the **counterfactual memorisation** (CM) metric (Feldman and Zhang, 2020):[2]

$$\texttt{CM}(x, y) = \underbrace{p_{\theta^{\text{IN}}}(y|x)}_{\text{training mem. (TM)}} - \underbrace{p_{\theta^{\text{OUT}}}(y|x)}_{\text{generalisation score (GS)}}$$

Here, θ^{IN} and θ^{OUT} represent models that have and have not seen (x, y) during training. The CM metric thus contrasts how a model performs on a training example to how a model *would have performed*, had the example not been in the training set; hence the 'counterfactual' nomenclature.

We compute an approximation of the TM, GS and CM metrics for 5M examples: 1M for 5 Indo-European language pairs (`En-Nl`, `-De`, `-It`, `-Fr`, `-Es`). We train 40 `transformer-base` (Vaswani et al., 2017) models from scratch per pair, on a parallel subcorpus constructed using OPUS data (Tiedemann and Thottingal, 2020). The 40 models have varying train and evaluation sets, such that TM, GS and CM can be computed for every example, averaging the quantities in the equation over outputs from multiple model instantiations. We put those 5M examples on a 'memorisation map', which we use to address the following sub-questions:

How do characteristics of datapoints relate to their position on the memorisation map? We compute 28 quantitative features and annotate a data subset manually using 7 additional features. We discuss how features such as source-target similarity, input and output length, token frequency and tokens' segmentation relate to the memorisation map. Figure 3 illustrates some key takeaways for different areas of the memorisation map, among which:

- As source–target overlap decreases, examples move down the diagonal. As a result, examples in the top right are near word-for-word translations, and misaligned examples are in the bottom left: truly misaligned examples are – even within a training regime with 1M examples – not memorised during training;
- Paraphrased and slightly inaccurate examples (in blue) can look alike according to a model;
- Examples with high CM scores tend to contain more infrequent words, be longer, and have higher BPE segmentation rates.

[2]This is a simplified representation of the metric. In practice, we replace probability with a geometric mean over target tokens' probabilities, to reduce length bias, and compute $p_{\theta^{\text{IN}}}$ and $p_{\theta^{\text{OUT}}}$ by averaging over results from various models.

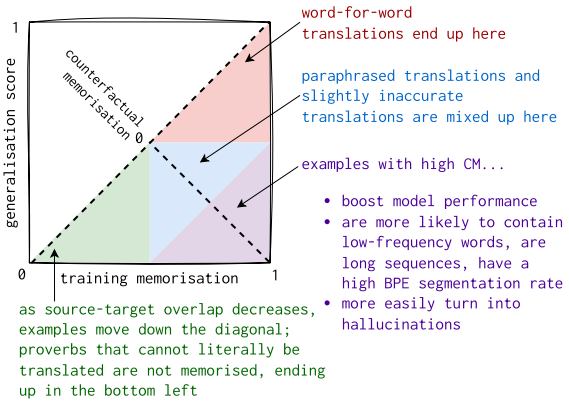


Figure 3: Illustrative summary of findings for different areas of the memorisation map. Counterfactual memorisation subtracts the y-coordinate from the x-coordinate. We detail some takeaways for areas of the map with high CM, low TM and low GS, high TM and high GS, and a similar TM and GS.

How do datapoints containing formulaic phrases stand out on the memorisation map? Although most experiments conducted here pertain to all training examples, we include an intermezzo to examine memorisation scores of source sequences that contain proverbs, idioms or non-compositional compounds. We would want these examples to be memorised *more*, yet, find that they are memorised *less*, emphasising that what is actually memorised is not necessarily what should be memorised.

Can we approximate memorisation metrics using datapoints' characteristics? Next, we use datapoints' characteristics to predict memorisation values with small regression models, allowing us to compare different language pairs and understand whether resource-intensive memorisation computation has cheaper approximates. We find that the regression models generalise cross-lingually, since characteristics' relation to memorisation is largely language-independent for our language pairs.

How does training on examples from different regions of the memorisation map change models' performance? Finally, we relate different parts of the map to the quality of NMT systems in terms of BLEU, COMET, targets' log-probability and models' hallucination tendency. Our results confirm that even in this real-world task, examples with high CM are most beneficial to model performance.

2.2 Conclusion and retrospective

En résumé, we contribute a valuable resource of memorisation scores, establish that memorisation

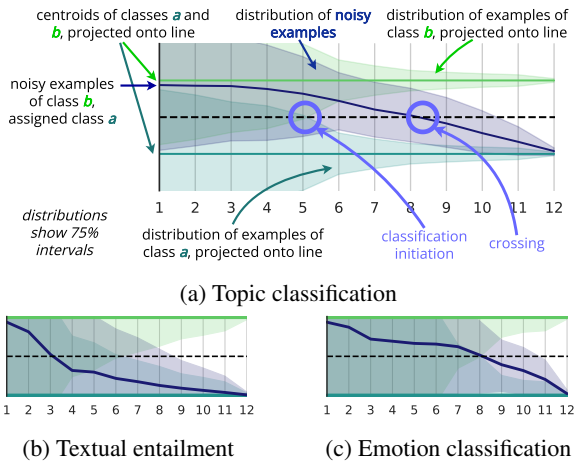
is not a mysterious phenomenon but a process that is predictable based on the features of data points and can positively benefit generalisation. While a valuable contribution, our work also had several limitations, among which were the computational expense of the experimental setup, and the focus on a set of five Indo-European language pairs. This was the consequence of a specific experimental design choice (parallel corpora across languages), which limits the generalisability of the findings.

Since the ideation of this project in 2022, the role of *large language models* (LLMs) in NMT has increased substantially. Because we train models from scratch – the default for NMT until ~2024 – it cannot assess how memorisation behaves during LLM fine-tuning or how pretraining might affect memorisation. Despite this, we consider the work a valuable contribution to the still-small body of research on CM in NLP (Raunak et al., 2021; Zheng and Jiang, 2022; Zhang et al., 2023). It was only the second to study CM at the scale of millions of examples (Zhang et al., 2023). Subsequent work has further examined memorised sequences in LLMs, echoing some of our findings (Prashanth et al., 2024). We hope that our per-datum memorisation estimates may serve as a benchmark in the future, for instance, benefiting the development of proxy metrics for CM.

It is also worth noting that, although this section takes the stance of CM being a generalisation benefit, not all types of memorisation are beneficial; memorisation is not a monolithic concept. In fact, in follow-up work (Dankers and Raunak, 2025), we show that knowledge distillation in NMT can increase extractive memorisation (a detrimental type of memorisation) and lead to more hallucinations.

3 Layer-based memorisation localisation (Dankers and Titov, 2024)

The previous section approached memorisation as a gradual phenomenon. Here, we instead focus on extreme memorisation of mislabelled examples to localise memorisation in LMs' layers. In *computer vision* (CV), studies tracing memorised, mislabelled examples often argue that lower layers learn generalisable features while deeper layers memorise (Cohen et al., 2018; Ansuini et al., 2019; Stephenson et al., 2021, i.a.). We refer to this as the *generalisation-first, memorisation-second* (GFMS) hypothesis. In NLP, localisation studies reach mixed conclusions when studying the memorisation of facts, idioms or verbatim sequences, pointing to the top (e.g. Dai et al., 2022; Zhao et al., 2024), early/middle (e.g. Meng et al., 2022), or lowest layers (e.g. Haviv et al., 2023; Stoehr et al., 2024). These conflicting findings may stem from varying experimental setups, and studying various memorisation types. Instead, we study the memorisation of mislabelled examples in NLP classification models to more directly compare with work from CV, putting the GFMS hypothesis to the test.



(a) Topic classification

(b) Textual entailment

(c) Emotion classification

Figure 4: Memorisation in action: an illustration of how memorised, mislabelled (noisy) examples move away from the centroid of their real class to their new class. Memorisation is gradual and non-local, but does shift to deeper layers for some tasks, such as topic classification.

3.1 Experiments and findings

We finetune four LMs (Devlin et al., 2019; Black et al., 2021; Zhang et al., 2022; Biderman et al., 2023) with a newly learned classification head across twelve datasets covering topic classification, sentiment analysis, hate speech detection, and generic NLU tasks (e.g. recognising textual entailment). We mislabel 15% of the examples and use them to study layer-based memorisation localisation, addressing the following sub-questions:

Can memorisation of mislabelled examples be localised to individual layers? Using four localisation methods – layer swapping, layer retraining, probing, and forgetting gradient analysis – we find that memorisation is not confined to single layers. Instead, multiple layers gradually shift mislabelled examples toward their assigned class, which is a process in which earlier layers are, relatively speaking, more important than later layers. To interpret this process, we introduce **centroid analysis** (Figure 4), which visualises how hidden representations of mislabelled examples change across layers.

How consistent is layer-based localisation across LMs and tasks? We observe subtle task differences linked to generalisation performance: when models generalise better, deeper layers contribute more to memorisation. Figure 4 illustrates this contrast. For instance, when comparing recognising textual entailment to topic classification, the latter generalises much better to test data than the former, while also having a 'crossing' of mislabelled examples in deeper layers in Figure 4. Comparing models with 12 and 24 layers shows that the lowest layers are not necessarily the most important in absolute terms, lowest is relative both with respect to model size and how 'deep' finetuning managed to change the pretrained model.

3.2 Conclusion and retrospective

Summarising, we localised memorisation by tracing mislabelled examples across layers. Memorisation was not confined to specific layers but emerged through cooperation across many layers, indicating that memorisation and generalisation are intertwined. However, layers contribute unequally: early layers play a larger role, as memorised examples begin to diverge there. These findings contradict the GFMS hypothesis. Because memorisation is distributed, local weight manipulations through editing or unlearning may alter behaviour without fully removing stored information.

Our work has several limitations. Mislabelled examples are only a proxy for real-world memorisation, and the localisation methods used are imperfect. Nonetheless, the agreement observed across LMs and methods suggests our conclusions are reliable. From a 2026 perspective, another limitation is our focus on 'traditional' fine-tuning with a task classification head. Yet, since the publication of our work, related studies have similarly argued that memorisation is distributed and intertwined with general language modelling abilities (Huang et al., 2024; Menta et al., 2025), suggesting our findings may extend beyond the specific setup studied.

Part II
(Non-)compositionality: a memorisation-generalisation case study

Let us now shift focus to (non-)compositionality. How does this reflect the tension between memorisation and generalisation, and how does memorisation of idioms affect models internally?

4 Evaluating (non-)compositional generalisation (Dankers et al., 2022a)

Compositionality plays an essential role in human language understanding, but whether neural networks exhibit this property has long been debated (e.g. Fodor and Pylyshyn, 1988; Smolensky, 1990; Marcus, 2003; Nefdt, 2020). Prior to 2022, studies of compositionality in NLP models mainly relied on synthetic datasets with simplified languages, where compositionality can be isolated and controlled (e.g. Lake and Baroni, 2018; Keysers et al., 2019; Hupkes et al., 2020; Kim and Linzen, 2020). These tests compute interpretations using a *local*, bottom-up notion of compositionality, ignoring that natural language contains exceptions such as idioms (see §1), which require more global sentence processing. For *natural* language, NLP systems must balance compositional and non-compositional processing. In this section, we analyse NMT outputs to explore this tension, contrasting compositional generalisation tests with the memorisation of non-compositional idioms. Before Dankers et al. (2022a), no datasets evaluated compositional generalisation in MT for models trained on natural language. We introduced new data to fill this gap and reformulated three theoretically grounded tests from Hupkes et al. (2020): systematicity, substitutivity, and overgeneralisation.

4.1 Experiments and findings

We train `transformer-base` (Vaswani et al., 2017) on a 1M, 8M, and 64M English-Dutch subset of the OPUS corpus (Tiedemann and Thottingal, 2020). We curate synthetic sentences in which we can control the lexical items and insert certain complex noun and verb phrases extracted from the OPUS data, yielding partially-natural, partially-synthetic evaluation data. We use the data and the models to answer the following research sub-questions:

How can we reformulate theoretically-grounded compositionality tests outside of toy task scenarios for NMT? We reimagine tests previously proposed by my co-authors and I (Hupkes et al., 2020) for English-Dutch data in an NMT setup:

- **Systematicity** evaluates the consistency of translations when recombining conjoined phrases with new phrases or when replacing words within a sentence (e.g. replacing "men" in "The girl sees that the men cry"). The recombinations are semantically unrelated and should not alter the translation when assuming

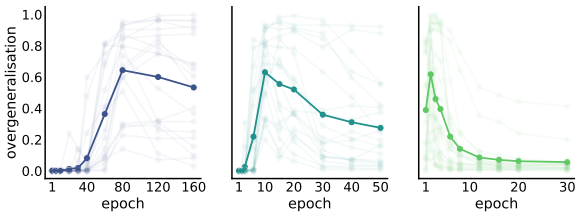


Figure 5: Overgeneralisation of idiomatic translations during training; an average is shown (bold) along with the trajectory of 20 individual idioms. From left to right, we show models trained on three training set sizes.

a locally compositional approach.

- **Substitutivity** evaluates translation consistency under synonym substitution, where two words in English map to the same Dutch word (e.g. ladybird/ladybug), using 20 synonym pairs. Since synonym substitution is meaning-preserving, the translations should not change.
- **Overgeneralisation** traces how 20 idioms, such as "out of the blue", are translated, assessing whether they are *overgeneralised* or *memorised*, based on the presence or absence of a literal translation of the idiom's keyword – i.e. translating "blue" as "blauw" would be overgeneralisation.

How compositional are NMT systems, and is the source of the errors natural language variation or model behaviour? Low systematicity and substitutivity consistency scores indicate that models often fail to behave compositionally under the strict local interpretation. We manually analyse 1800 inconsistent translation pairs, identifying that some inconsistencies reflect natural variation in language, but that the majority cannot reasonably be traced to linguistic ambiguity, underscoring NMT models' general volatility. This erratic behaviour highlights a lack of default reasoning, which can be problematic or even harmful in some cases, especially if faithfulness (Parthasarathi et al., 2021) or consistency is important to the end user.

How do NMT systems acquire non-compositional translations of idioms, and how does this align with generalisation performance? The third test demonstrates that models acquire idiomatic translations in two phases, as shown in Figure 5: early in training, the models learn to overgeneralise word-for-word translations, and later, they start to memorise paraphrases. Models' convergence based on memorisation does not appear to align with the other evaluation metrics; especially for the models trained on 1M and 8M data, more training would be needed to achieve memorisation. Interestingly, these models are simultaneously *not compositional enough* (as per the systematicity and substitutivity tests) and *too compositional* (as per the overgeneralisation test).

4.2 Conclusion and retrospective

Research on compositional generalisation often relies on artificial tasks that assume strictly local interpretations of compositionality. We argued that such interpretations overlook important aspects of natural language and proposed evaluating compositionality in NMT systems trained on natural data. We reformulated three compositionality tests showing that models simultaneously struggle with compositional generalisation and adequate memorisation of idioms. These findings highlight the difficulty of evaluating compositionality in natural language, where meaning composition is less clear-cut than in synthetic datasets. Following Baggio (2021), we suggest that human-like language use likely requires models to support both behaviours.

Our work also has limitations. Although we argue that compositional generalisation should ideally be evaluated using fully natural data, we rely on partially synthetic tests and do not propose direct solutions for improving compositional generalisation or non-compositional memorisation. Subsequent work, though, has leveraged our findings for actionable training techniques such as consistency regularisation (Yin et al., 2023) and novel dropout schemes (Niculae and Monz, 2023). Other studies have adapted our three tests to different languages, domains, and modalities (e.g. Liu, 2022; Li et al., 2024a; Liao et al., 2023; Kumon et al., 2024; Moisio et al., 2023). More broadly, we not only highlighted models' limitations but also explicitly encouraged the community to rethink how compositional generalisation is evaluated, rather than removing natural language variation for convenience. This call has been widely echoed (e.g. Zheng and Lapata, 2023; Sun et al., 2023; Chia, 2024; Chia et al., 2024; Fodor et al., 2025) and may be the most influential outcome of this research.

5 Mechanisms for idiomatic translations (Dankers et al., 2022b)

Having introduced the tension between compositional and non-compositional processing and framed idiom acquisition as a two-step process,

we now examine how pretrained models perform idiom translation. Idioms have long challenged NLP (e.g. Sag et al., 2002; Rayson et al., 2010; Shwartz and Dagan, 2019), particularly NMT systems (e.g. Barreiro et al., 2013; Isabelle et al., 2017; Constant et al., 2017; Avramidis et al., 2019). Not all *potentially idiomatic expressions* (PIEs) are figurative – e.g. consider "When I kicked the bucket, it fell over" – so correct translation depends on the context.[3] NMT systems must therefore learn to disambiguate usage, memorise the appropriate paraphrase, and generate it during decoding. Until the presentation of our work, the neural mechanisms underlying idiomatic translation remained poorly understood. Earlier work mainly examined how transformer-based LMs represent idioms (e.g. García et al., 2021a,b), but LMs merely need to detect figurativeness; they are not trained to explicate the idiomatic meaning. We present the first large-scale analysis of how transformers translate idioms, investigating whether models paraphrase or translate them word for word, and analysing their effects on self- and cross-attention as well as encoder hidden states.

5.1 Experiments and findings

We analyse `transformer-base` models (Vaswani et al., 2017) pretrained by Tiedemann and Thottingal (2020) for seven Indo-European language pairs (`En-Nl`, `-De`, `-Sv`, `-Da`, `-It`, `-Fr`, `-Es`) by comparing literal and figurative occurrences of PIEs. We address the following research sub-questions:

How can we perform analyses of NMT idiom processing at scale? Large-scale analyses of idiom translations suffer from a lack of parallel corpora (Fadaee et al., 2018). We therefore perform our analyses using data from the monolingual MAGPIE corpus (Haagsma et al., 2020), for which we devise a heuristic translation annotation method (similar to §4, but at a larger scale). We extract translations from our seven models and use a list of literal translations of idiom keywords to distinguish paraphrases from word-for-word translations. To validate the heuristic, we conducted human data annotation. Figurative PIEs should generally not be translated word for word due to their non-compositional meaning; however, only 20.7% of translations were paraphrased by the models. This

[3]Up to this point, we referred to idioms, but since this section considers both literal and figurative occurrences, we mainly use the term PIEs.

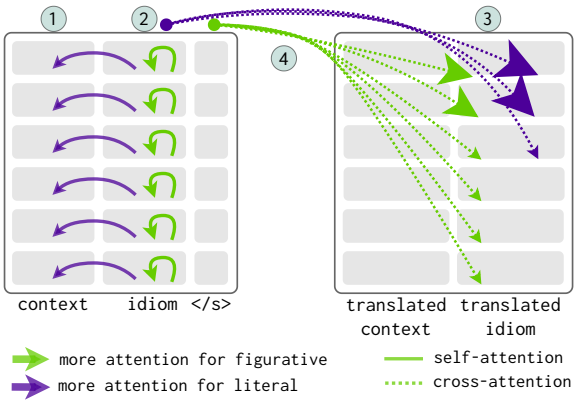


Figure 6: When comparing figurative, paraphrased PIEs to literal PIEs translated word for word, we find (1) less attention from PIE to context, (2) more attention within the PIE, (3) less cross-attention between PIE and its paraphrase, (4) more cross-attention from the paraphrase to `</s>`. The magnitude of the arrowhead indicates the effect size.

echoes findings from §4 that transformers can be *too* compositional, likely due to insufficient memorisation of idioms (§2). We further show that idiom frequency in the training corpus partially explain this.

How does idiomaticity and the paraphrasing of non-compositional idioms affect attention patterns and hidden representations? In the encoder, figurative PIEs are attended to more strongly as a single lexical unit than literal instances and interact less with the surrounding context. This aligns with prior work (e.g. Zaninello and Birch, 2020) showing that encoding idioms as single units improves translation. While paraphrasing PIEs, the decoder attends less to source tokens and more to the end-of-sentence token, temporarily detaching from the encoder input. Figure 6 visually summarises these findings. Hidden-state analysis confirms that these attention shifts affect the residual stream. The effects occur across layers: PIE representations gradually become more distinct, with figurative PIEs standing out from the first layer onward.

How do encoder-internal interventions affect non-compositional translations? We, lastly, intervene in the encoding of the PIEs using amnesic probing (Elazar et al., 2021), demonstrating that one can easily change non-compositional translations into compositional ones, underscoring that recalling memorised PIE paraphrases is a brittle process. When doing so, the transformer's self-attention changes such that the PIE components are grouped *less*, strengthening our previous findings.

Figure 7: Example of an overly compositional translation in Frisian ("chickens came home to roost" is included literally, but is not a Frisian idiom). Retrieved on March 15, 2026, from `GPT-5.3`.

5.2 Conclusion and retrospective

Summarising, we showed that idioms are often translated too compositionally and presented analyses of transformer mechanisms for paraphrasing idioms, pointing to a grouping mechanism in self-attention as key for treating idioms as single, non-compositional units. Our work had several limitations, most notably the focus on high-resource languages (due to the lack of high-quality idiom translations for low-resource languages) and the heuristic labelling of idiom translations. Nonetheless, it was among the first influential interpretability analyses of idiom translation. Closely related work includes Baziotis et al. (2023), Haviv et al. (2023), and Lim et al. (2024).

Although the lack of parallel idiom corpora remains a challenge, several multilingual idiom translation datasets have now been introduced (e.g. Amrhein et al., 2022; Stap et al., 2024; Lee et al., 2025), along with methods for improving idiom translations (e.g. Santing et al., 2022; Li et al., 2024b; Liu et al., 2023; Donthi et al., 2025). The largest translation quality gains since 2022, however, have come from scaling pretraining corpora for ever-growing LLMs. And yet, for the most powerful (non-)commercial translation systems and LLMs, researchers continued to echo our findings of overly compositional translations for a range of languages, such as German, Spanish and Japanese (Ferrando et al., 2023), Arabic (Obeidat et al., 2024), Urdu (Basit et al., 2024) and Indonesian (Dewayanti and Margana, 2024). In 2026, it still takes mere minutes to identify idiom translation failures for the most powerful models, particularly for low-resource languages (see Figure 7). Idiom translation has vastly improved, but it is not quite there yet!

6 Conclusion

Across the different sections, we investigated memorisation in neural models, its relationship to generalisation, and its connection to (non-)compositionality. Seven main lessons emerged with respect to the research questions I introduced in §1.

What characterises memorised examples? We find that *(1) memorisation is predictable rather than mysterious*. In §2, we introduced memorisation scores for 5M MT examples (with *Counterfactual Memorisation* being the core focus) showing that memorisation exists along a continuum. Much of the variation in these scores can be explained by surface-level features such as source-target overlap, and these patterns generalised across five language pairs. For idioms, a characteristic influencing memorisation of paraphrased translations was the frequency in the training corpus (§5). Second, *(2) what requires memorisation is not necessarily what models memorise under standard training regimes*. NMT systems often fail to memorise idioms and instead translate them compositionally (§2,4,5).

Which model-internal mechanisms enable memorisation? *(3) Memorisation is distributed across layers rather than localised*. In §3, localisation experiments on four transformer LMs across twelve tasks showed that memorisation of mislabelled examples emerges through cooperation across layers. Hidden representations gradually shift toward memorised labels rather than changing in a single layer, and deeper layers do not play a uniquely dominant role. Furthermore, *(4) idiom memorisation in translation involves grouping on the source side and reduced reliance on the encoder during decoding*. Attention analyses in §5 revealed that paraphrased idioms exhibit increased internal attention and reduced interaction with surrounding context, suggesting that they are processed as single units. During decoding, attention shifts away from idiom tokens toward the EOS token.

To what extent are memorisation and generalisation at odds with one another? *(5) Memorising atypical examples can support generalisation*. Experiments in §2 show that examples with higher CM scores benefit models' translation performance most, likely because examples with high CM are not merely noise, but representative of natural language variation in translations. Next, *(6) idiom acquisition follows a multi-phase process*. In §4,

tracing translations during training revealed an initial overgeneralisation phase followed by memorisation. For frequent idioms, models eventually produce memorised paraphrases, but many idioms remain in the overgeneralisation phase (§5), yielding overly compositional translations. Finally, *(7) transformers do not process language in a locally compositional manner*. By adapting synthetic compositional generalisation tests to natural-language MT data (§4), we identified that when we expect models to be locally compositional, they can actually be very volatile and inconsistent in their translations. This underscores a paradox: models are simultaneously not *compositional* and not *non-compositional* enough.

A final retrospective and outlook In summary, we learnt that transformer models memorise substantial aspects of their training data, which can support generalisation in natural language tasks. However, they often fail to memorise the types of formulaic expressions that require it, such as idioms. At the same time, transformers display both insufficient and excessive compositionality: non-local processing supports memorisation of idioms, yet harms compositional generalisation. Memorisation mechanisms emerge naturally but are distributed across layers and remain insufficiently adapted to natural language's formulaic nature.

Based on my findings, I propose several directions for future work. First, evaluations of compositional generalisation should not avoid non-compositional phenomena in natural language; methods that improve generalisation should also consider their implications for proverbs and idioms. Second, memorisation should be studied more holistically by focusing on memorisation circuits rather than individual neurons or layers. Accordingly, model editing and unlearning should move beyond layer-local approaches if the goal is true erasure of information. Third, memorisation can be beneficial and should be more explicitly incorporated into LLM design. Finally, as models are increasingly trained on their own generations, we risk losing subtle linguistic phenomena – such as idiomatic and proverbial expressions – that may become overgeneralised across languages. Which parts of natural language are we losing if transformer's own predictions become a part of that language? This warrants carefully crafted investigations, such as measuring the prominence of formulaic language in real and synthesised corpora.

Acknowledgments

While I refer the reader to my dissertation for the full acknowledgments, I'd like to repeat that I'm particularly grateful to my supervisor Ivan Titov and research mentor Dieuwke Hupkes for their support, expertise, guidance, insights, and enthusiasm throughout my PhD.

I, furthermore, thank Marius Mosbach and Cesare Spinoso-Di Piano for their feedback on the summary presented to you in this document.

Throughout the PhD I (Verna Dankers) was 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.

AI Assistant Usage

ChatGPT 5.3/5.4 was used to assist with the writing, primarily through shortening and proofreading. The AI assistant was not used to generate experimental results, to make scientific claims, or to determine conclusions. All experiments from the original papers, and all text therein, were produced without any assistance from LLMs.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. Aces: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513.

Jacob Andreas. 2018. Measuring compositionality in representation learning. In *International Conference on Learning Representations*.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32:6111–6122.

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.

Giosuè Baggio. 2021. Compositionality in a parallel architecture for language processing. *Cognitive Science*, 45(5):e12949.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, Fernando Batista, and 1 others. 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.

Abdul Basit, Abdul Hameed Azeemi, and Agha Ali Raza. 2024. Challenges in Urdu machine translation. In *Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 44–49.

Christos Baziotis, Prashant Mathur, and Eva Hasler. 2023. Automatic evaluation and analysis of idioms in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3682–3700.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327.

Zheng Chia. 2024. *Exploring Optimal Settings for Machine Translation of Irony with Application to Multilingual Irony Detection*. Phd thesis, Kitami Institute of Technology.

Zheng Lin Chia, Michal Ptaszynski, Marzena Karpinska, Juuso Eronen, and Fumito Masui. 2024. Initial exploration into sarcasm and irony through machine translation. *Natural Language Processing Journal*, 9:100106.

Gilad Cohen, Guillermo Sapiro, and Raja Giryes. 2018. DNN or k-NN: That is the generalize vs. memorize question. *arXiv preprint arXiv:1805.06822*.

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.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022a. 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 (Volume 1: Long Papers)*, pages 4154–4175.

Verna Dankers and Christopher Lucas. 2023. Non-compositionality in sentiment: New data and analyses. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5150–5162.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022b. Can transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626.

Verna Dankers and Vikas Raunak. 2025. Memorization inheritance in sequence-level knowledge distillation for neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 760–774.

Verna Dankers and Ivan Titov. 2022. Recursive neural networks with bottlenecks diagnose (non-)compositionality. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4361–4378.

Verna Dankers and Ivan Titov. 2024. Generalisation first, memorisation second? Memorisation localisation for natural language classification tasks. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14348–14366.

Verna Dankers, Ivan Titov, and Dieuwke Hupkes. 2023. Memorisation cartography: Mapping out the memorisation-generalisation continuum in neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8323–8343.

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.

Desakh Putu Setyalika Putri Dewayanti and Margana Margana. 2024. The impact of contextual understanding on neural machine translation accuracy: A case study of Indonesian cultural idioms in English translation. *Englisia: Journal of Language, Education, and Humanities*, 12(1):223–236.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181.

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)*.

Vitaly Feldman. 2020. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2881–2891.

Javier Ferrando, Matthias Sperber, Hendra Setiawan, Dominic Telaar, and Saša Hasan. 2023. Automating behavioral testing in machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1014–1030.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2025. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

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.

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.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047.

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10711–10732.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intellgence Research*, 67:757–795.

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.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, and 1 others. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *The Seventh International Conference on Learning Representations*.

Najoung Kim and Tal Linzen. 2020. COGS: a compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105.

Tomasz Korbak, Julian Zubek, and Joanna Rączaszek-Leonardi. 2020. Measuring non-trivial compositionality in emergent communication. In *NeurIPS 2020 workshop on Emergent Communication.*

Ryoma Kumon, Daiki Matsuoka, and Hitomi Yanaka. 2024. Evaluating structural generalization in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13220–13239.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2873–2882. PMLR.

Minjae Lee, Youngbin Noh, and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646.

Chuanhao Li, Zhen Li, Chenchen Jing, Yuwei Wu, Mingliang Zhai, and Yunde Jia. 2024a. Compositional substitutivity of visual reasoning for visual question answering. In *European Conference on Computer Vision*, pages 143–160. Springer.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024b. Translate meanings, not just words: IdiomKB's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Weiduo Liao, Ying Wei, Mingchen Jiang, Qingfu Zhang, and Hisao Ishibuchi. 2023. Does continual learning meet compositionality? New benchmarks and an evaluation framework. *Advances in Neural Information Processing Systems*, 36:33499–33513.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111.

Yutong Liu. 2022. Compositional generalization in machine translation for low-resource languages. Master's thesis, University of Edinburgh.

Gary F Marcus. 2003. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Tarun Ram Menta, Susmit Agrawal, and Chirag Agarwal. 2025. Analyzing memorization in large language models through the lens of model attribution. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10661–10689.

Anssi Moisio, Mathias Creutz, and Mikko Kurimo. 2023. On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 204–213.

Ryan M Nefdt. 2020. A puzzle concerning compositionality in machines. *Minds & Machines*, 30(1).

Vlad Niculae and Christof Monz. 2023. Joint dropout: Improving generalizability in low-resource neural machine translation through phrase pair variables. *MT Summit 2023*, page 12.

Mohammed M Obeidat, Ahmad S Haider, Sausan Abu Tair, and Yousef Sahari. 2024. Analyzing the performance of Gemini, ChatGPT, and Google Translate in rendering English idioms into Arabic. *FWU Journal of Social Sciences*, 18(4).

Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. 2021. Sometimes we want ungrammatical translations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3205–3227.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir SV, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, and 1 others. 2024. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*.

Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, and Aline Villavicencio. 2016. How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 156–161.

Vikas Raunak and Arul Menezes. 2022. Finding memo: Extractive memorization in constrained sequence generation tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5153–5162.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begona Villada Moirón. 2010. Multiword expressions: Hard going or plain sailing? *Language Resources and Evaluation*, 44:1–5.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002, Proceedings*, pages 1–15. Springer.

Lukas Santing, Ryan Sijstermans, Giacomo Anerdi, Pedro Jeuris, Marijn ten Thij, and Riza Batista-Navarro. 2022. Food for thought: How can we exploit contextual embeddings in the translation of idiomatic expressions? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 100–110.

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.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6189–6206.

Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and Sue Yeon Chung. 2021. On the geometry of generalization and memorization in deep neural networks. In *The Ninth International Conference on Learning Representations*.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*.

Kaiser Sun, Adina Williams, and Dieuwke Hupkes. 2023. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 274–293.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

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.

Lena Voita. 2024. Analysis methods for natural language processing.

Alison Wray. 2002. *Formulaic language and the lexicon.* ERIC.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, Jie Zhou, and Yue Zhang. 2023. Consistency regularization training for compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1294–1308.

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.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2088–2102.

Hao Zheng and Mirella Lapata. 2023. Real-world compositional generalization with disentangled sequence-to-sequence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1711–1725.

Xiaosen Zheng and Jing Jiang. 2022. An empirical study of memorization in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6265–6278.

A Advice from beyond the PhD

① Don't stress but plan ahead Questions I often received towards the end of the PhD were along the lines of "How do I ensure that my thesis has a cohesive storyline?" and "Did you plan for this story from the start?". Answering them is non-trivial: in a fast-paced field like NLP, planning three to six years ahead for a *top-down* approach to PhD research is nearly impossible, and many a thesis is constructed in a *bottom-up* manner, where the candidate bundles their papers towards the end, determining the storyline somewhat post-hoc. This works well for many, so do not worry too much if you are currently in a situation in which you can't completely see the dissertation's forest through your papers' trees yet. However, I personally very much enjoy the fact that my dissertation has clear, connected themes.

I believe that crucial reasons for why this is the case are (a) that I entered the PhD with a strong interest in (non-)compositionality and generalisation and didn't stray too far from that through the years, while having a flexibility regarding projects to conduct along the way, and (b) that I planned my projects ahead of time, supported by my university's administrative structures. During three intermediate PhD evaluations, I planned ahead. During evaluation 1, I sketched out research directions for the next year, during evaluation 2, I discussed how the thesis's storyline was starting to emerge, and what was missing, and during evaluation 3, I discussed a concrete thesis outline. This does not mean that I stuck exactly to what was planned (see lesson ②, below), but it was a massive help. If you're a student, and your university/supervisor does not enforce such evaluation and planning checkpoints, be proactive and schedule them yourself, at least once a year.

For me, the right approach was not working bottom-up, not working top-down, but adopting a *hybrid* approach, where most meetings are low-level, bottom-up project meetings, but some take the helicopter view of the PhD as a marathon, to study whether I was on track, and was headed in the right direction.

② Curate the thesis's storyline I can't say with certainty that I did it well, but I attempted to curate a very specific storyline in the dissertation. Three things I did to achieve that, are the following:

Selecting what to include. Not everything one does during a PhD has to be included in the thesis, and even if you include all first-author papers, you can often write multiple stories using the same papers. If you've read the summary of my thesis above, you will have noticed that it contains two primary parts, circled around 'memorisation vs generalisation' and 'compositionality vs non-compositionality'. Planning for the thesis writing, I originally thought I'd include a Part III – based on Dankers and Titov (2022); Dankers and Lucas (2023), as summarised in Appendix B – but I followed the advice of my UoE annual review board to omit it. More isn't always better.

Deciding where to include what. In my case (but perhaps not yours) the papers didn't have to be included in a chronological order, and I opted for putting the most recent two first to curate a narrative in which the second two papers (on (non-)compositionality) are considered a case study of the more general memorisation–generalisation paradox.

Going beyond the original papers. Every experimental chapter of my thesis contains experiments that were not in the original paper, either because I felt something was missing that could have made the conclusions stronger, or because seeing the individual papers in the light of the overall dissertation made me realise that certain experiments could provide inter-chapter connections.

③ Plan for reproducibility By the time the thesis comes around you might be getting back into 4-year-old codebases, to see whether you can recreate some graphs with old data – either just because the thesis would look nicer with updated graphs, or because you actually want to extend the experiments, like I did. At that point, you might realise that not everything is reproducible. Even with the greatest `README.md` out there, packages will change, clusters will change, and checkpoints you thought you had stored will no longer be there because the cluster admin decided a clean-up was in order. When you're now only one year into the PhD, I know you're likely not thinking of the dissertation, and you may want to move on from a project as soon as it is submitted, but please invest time into the reproducibility aspect. First and foremost because of others that might want to build upon your work, and second because you yourself may want to re-run those exact experiments. Think ahead: "Which model checkpoints would be crucial if I wanted to test a few more hypotheses for the thesis?", or

"What data would I need to store if I wanted to regenerate the figures, or run an additional statistical significance test on the results?", or "Should I ask my internship manager for approval for exporting these models, since I won't be an intern at the time of graduation?".

I wish I had curated a PhD-thesis-ready version of all of my papers on a dedicated hard drive, with the most interesting models, and the data required to regenerate tables and graphs. Although it was there for most of the chapters, some of it needed to be regenerated.

④ **Give your dataset a name** A very silly mistake I realised I made was that in Dankers et al. (2022a) I produced a compositional generalisation evaluation dataset that never received a name. Therefore, various papers referred to it under names they had come up with. The consensus was 'OPUS En-Nl', but we clearly should have assigned the evaluation set a name. Lesson learnt!

⑤ **Write a retrospective** Inspired by Voita (2024)'s sections entitled "Implications: View from the future", I wrote my own "Retrospective and outlook" sections, which were also highly recommended by my phenomenal supervisor Prof Titov. Taking my own work, and reflecting on (a) what impact this specific paper has had and (b) how the field has changed since this paper saw the light of day, was very informative for myself, and according to my thesis examiners. I can safely say they are my favourite thesis sections.

B The 'Lost' Part III: Quantifying (non-)compositionality

There are two papers that I initially planned to include in the thesis, but did not, in the end (Dankers and Titov, 2022; Dankers and Lucas, 2023). These articles focus on the notion of (non-)compositionality: Across sentences and even across subphrases of one sentence, there is variation in terms of how compositional phrases are. While there is a wealth of knowledge about how very specific types of non-compositional phenomena behave, quantifying the compositionality of phrases or sentences in general is an open and ill-defined problem. Both of these articles address this open problem from different angles by using either model-computed or human-derived quantifications.

Recursive model-based metric (Dankers and Titov, 2022) Gaining a better understanding of the challenges that the (non-)compositionality of natural language presents to neural models, requires metrics for quantifying that (non-)compositionality. While this had previously been investigated on a small scale, such as for idioms (Ramisch et al., 2016), for natural language bigrams (Andreas, 2018), or for artificial languages in the context of communication games (Korbak et al., 2020), the metrics proposed for those applications did not easily scale to natural language sentences.

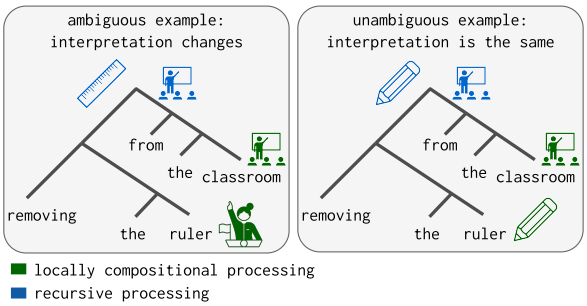


Figure 8: When processing this phrase, "the ruler" is interpreted differently when comparing recursive processing with local processing. We enforce local processing by equipping models with bottlenecks, and our **bottleneck compositionality metric (BCM)** then compares inputs' representations before and after compression through the bottleneck.

In this work, we extended one such metric, namely the *Tree Reconstruction Error* (TRE) by Andreas (2018), that expresses the distance between a model's representation of an input and a strictly compositional reconstruction of that representation. To do so, we used recursive neural networks, namely Tree-LSTMs (Tai et al., 2015), to process inputs according to their syntactic structure. We augmented Tree-LSTMs with bottlenecks to compute the meaning of an input with respect to a certain task in a more locally compositional manner. We used these models to distinguish more compositional examples from less compositional ones in a *bottleneck compositionality metric* (BCM). BCM compares a regular model trained to perform a task to a model augmented with bottlenecks. Under the assumption that non-compositional processing of an input requires more complex meaning representations, the bottlenecks will hinder examples for which the model finds a non-compositional solution most, as is illustrated in Figure 8. As such, the difference between a model without the bottlenecks and one with the bottlenecks acts as a metric.

We experimented with three types of bottlenecks,

namely a *deep variational information bottleneck* (DVIB) (Alemi et al., 2017), compressing representations through increased dropout or simply using smaller hidden dimensionalities. As a proof of concept, the BCM was applied in a controlled environment where non-compositional examples were manually introduced, by taking arithmetic expressions and making one vocabulary item ambiguous. Afterwards, we applied the BCM to the real-world example of sentiment analysis using the *Stanford Sentiment Treebank* (SST) dataset (Socher et al., 2013). For both tasks, we illustrated that compression through a bottleneck encourages local processing, and showed that the bottleneck can act as a metric distinguishing compositional from less compositional samples. We, furthermore, used the compositionality judgments for the SST data to demonstrate that (i) in a training data scarce scenario, compositional training examples yield models that generalise better to test data, and (ii) that when the test set contains non-compositional examples, performance is substantially lower compared to a test set of compositional examples.

Using human annotations for compositionality judgments (Dankers and Lucas, 2023) In this work, we used a different approach and focused on human data annotations to obtain quantifications of natural language phrases' (non-)compositionality. For such phrases, their meaning is often more than 'just' the compositional sum of their parts. In the context of sentiment analysis, the 'meaning' of a phrase is its sentiment, and even though sentiment computations are largely compositional, there are still exceptional patterns. We selected the task of sentiment analysis as a testbed for obtaining non-compositionality ratings for phrases because of this rather straightforward interpretation of 'meaning', which is much less well defined in other tasks or when obtaining task-generic compositionality judgments.

We first designed a protocol to obtain non-compositionality judgments based on human-annotated sentiment. Our methodology uses phrases from the SST dataset (Socher et al., 2013) and contrasts the sentiment of a phrase with that of control stimuli, in which one of two subphrases has been replaced. Phrases whose annotated sentiment deviates from what is expected are considered less compositional. This approach, along with the example of the non-compositional phrase "all the excitement of eating oatmeal", is depicted

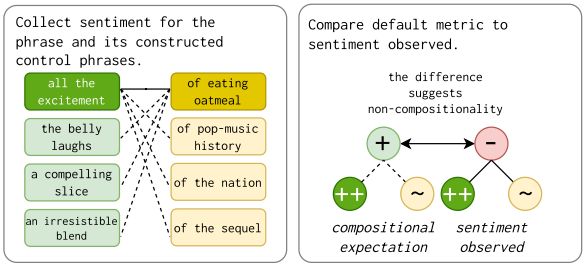


Figure 9: Illustration of how the non-compositionality ratings are obtained in NONCOMPSST: we contrast a sentiment's phrase to the sentiment of control phrases.

in Figure 9. We developed a resource of ratings for 259 phrases (dubbed NONCOMPSST) through sentiment annotations from 147 participants total, who provided us with more than 10,000 annotations via Prolific. Secondly, we presented an analysis of that resource, emphasising the higher non-compositionality ratings for figurative language, and use NONCOMPSST to evaluate computational models for sentiment analysis, focusing on conventional pretrained models such as ROBERTA, as well as models pretrained on sentiment-laden data, and models finetuned for sentiment analysis. Performance on conventional SST test data was markedly higher compared to performance on NONCOMPSST, underscoring that non-compositional phrases challenge models more. We suggest that NONCOMPSST can complement existing evaluation protocols for sentiment analysis models.