The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study

Verna Dankers
ILCC, University of Edinburgh
vernadankers@gmail.com

Elia Bruni
University of Osnabrück
elia.bruni@gmail.com

Dieuwke Hupkes
Facebook AI Research
dieuwkehupkes@fb.com

Abstract

Obtaining human-like performance in NLP is often argued to require compositional generalisation. Whether neural networks exhibit this ability is usually studied by training models on highly compositional synthetic data. However, compositionality in natural language is much more complex than the rigid, arithmetic-like version such data adheres to, and artificial compositionality tests thus do not allow us to determine how neural models deal with more realistic forms of compositionality. In this work, we re-instantiate three compositionality tests from the literature and reformulate them for *neural machine translation* (NMT). Our results highlight that: i) unfavourably, models trained on *more* data are more compositional; ii) models are sometimes less compositional than expected, but sometimes more, exemplifying that different *levels* of compositionality are required, and models are not always able to modulate between them correctly; iii) some of the non-compositional behaviours are mistakes, whereas others reflect the natural variation in data. Apart from an empirical study, our work is a call to action: we should rethink the evaluation of compositionality in neural networks and develop benchmarks using *real* data to evaluate compositionality on natural language, where composing meaning is not as straightforward as doing the math.[1]

1 Introduction

Although the successes of deep neural networks in *natural language processing* (NLP) are astounding and undeniable, they are still regularly criticised for lacking the powerful generalisation capacities that characterise human intelligence. A frequently mentioned concept in such critiques is *compositionality*: the ability to build up the meaning of a complex expression by combining the meanings of its parts (e.g. Partee, 1984). Compositionality is assumed to play an essential role in how humans understand language, but whether neural networks also exhibit this property has since long been a topic of vivid debate (e.g. Fodor and Pylyshyn, 1988; Smolensky, 1990; Marcus, 2003; Nefdt, 2020).

Studies about the compositional abilities of neural networks consider almost exclusively models trained on synthetic datasets, in which compositionality can be ensured and isolated (e.g. Lake and Baroni, 2018; Hupkes et al., 2020).[2] In such tests, the interpretation of expressions is computed completely *locally*: every subpart is evaluated independently – without taking into account any external context – and the meaning of the whole expression is then formed by combining the meanings of its parts in a bottom-up fashion. This protocol matches the type of compositionality observed in arithmetic: the meaning of $(3 + 5)$ is always 8, independent of the context it occurs in.

However, as exemplified by the sub-par performance of symbolic models that allow only strict, local protocols, compositionality in natural domains is far more intricate than this rigid, arithmetic-like variant of compositionality. Natural language seems very compositional, but at the same time, it is riddled with cases that are difficult to interpret with a strictly local interpretation of compositionality. Sometimes, the meaning of an expression does not derive from its parts (e.g. for idioms), but the parts themselves are used compositionally in other contexts. In other cases, the meaning of an expression does depend on its parts in a compositional way, but arriving at this meaning requires a more *global* approach because the meanings of the parts need to be disambiguated by information from elsewhere. For instance, consider the meaning of homonyms ("these dates are perfect for our dish/wedding"), potentially idiomatic expressions ("the child kicked the bucket off the pavement"),

[1] The data and code are available at https://github.com/i-machine-think/compositionality_paradox_mt. We present details concerning reproducibility in Appendix E.

[2] Apart from Raunak et al. (2019), work on compositionality and 'natural' language considers highly structured subsets of language (e.g. Kim and Linzen, 2020; Keysers et al., 2019).

or scope ambiguities ("every human likes a cat"). This paradoxical tension between local and global forms of compositionality inspired many debates on the compositionality of natural language. Likewise, it impacts the evaluation of compositionality in NLP models. On the one hand, local compositionality seems necessary for robust and reliable generalisation. Yet, at the same time, global compositionality is needed to appropriately address the full complexity of language, which makes evaluating compositionality of state-of-the-art models 'in the wild' a complicated endeavour.

In this work, we face this challenge head-on. We concentrate on the domain of *neural machine translation* (NMT), which is paradigmatically close to the tasks typically considered for compositionality tests, where the target represents the 'meaning' of the input.[3] Furthermore, MT is an important domain of NLP, for which compositional generalisation is important to produce more robust translations and train adequate models for low-resource languages (see, e.g. Chaabouni et al., 2021). As an added advantage, compositionality is traditionally well studied and motivated for MT (Rosetta, 1994; Janssen and Partee, 1997; Janssen, 1998).

We reformulate three theoretically grounded tests from Hupkes et al. (2020): *systematicity*, *substitutivity* and *overgeneralisation*. Since accuracy – commonly used in artificial compositionality tests – is not a suitable evaluation metric for MT, we base our evaluations on the extent to which models behave *consistently*, rather than correctly. In our tests for systematicity and substitutivity, we consider whether processing is *local*; in our overgeneralisation test, we consider how models treat idioms that are assumed to require *global* processing.

Our results indicate that models often do not behave compositionally under the local interpretation, but exhibit behaviour that is *too local* in other cases. In other words, models have the ability to process phrases both locally and globally but do not always correctly modulate between them. We further show that some inconsistencies reflect variation in natural language, whereas others are true *compositional mistakes*, exemplifying the need for both local and global compositionality as well as illustrating the need for tests that encompass them both.

With our study, we contribute to ongoing questions about the compositional abilities of neural networks, and we provide nuance to the nature of this question when natural language is concerned: how local should the compositionality of models for natural language actually be? Aside from an empirical study, our work is also a call to action: we should rethink the evaluation of compositionality in neural networks and develop benchmarks using *real* data to evaluate compositionality on natural language, where composing meaning is not as straightforward as doing the math.

2 Local and global compositionality

Tests for compositional generalisation in neural networks typically assume an arithmetic-like version of compositionality, in which meaning can be computed bottom up. The compositions require only local information – they are context independent and unambiguous: "walk twice after jump thrice" (a fragment from SCAN by Lake and Baroni, 2018) is evaluated similarly to $(2 + 1) \times (4 - 5)$. In MT, this type of compositionality would imply that a change in a word or phrase should affect only the translation of that word or phrase, or at most the smallest constituent it is a part of. For instance, the translation of "the girl" should not change depending on the verb phrase that follows it, and in the translation of a conjunction of two sentences, making a change in the first conjunct should not change the translation of the second. While translating in such a local way seems robust and productive, it is not always realistic – e.g. consider the translation of "dates" in "She hated bananas and she liked dates".

In linguistics and philosophy of language, the *level* of compositionality has been widely discussed, which led to a variety of definitions. One of the most well-known ones is from Partee (1984):

> "The meaning of a compound expression is a function of the meanings of its parts and of the way they are syntactically combined."[4]

This definition hardly places restrictions on the relationship between expressions and their parts. The type of function that relates them is unspecified and could take into account the global syntactic structure or external arguments, and the parts' meanings can depend on global information. Partee's definition is therefore called *weak*, *global*, or *open* compositionality (Szabó, 2012; García-Ramírez, 2019). When, instead, the meaning of a compound depends only on the meanings of its largest parts, regardless of their internal structure (similar to arithmetic), that is *strong*, *local* or *closed*

[3] E.g. SCAN's inputs are instructions ("walk twice") with executions as outputs ("walk walk") (Lake and Baroni, 2018).

[4] This straightforwardly extends to translation, by replacing *meaning* with *translation* (Rosetta, 1994).

n	Template
1	The N_{people} V the N^{sl}_{people} .
2	The N_{people} Adv V the N^{sl}_{people} .
3	The N_{people} P the $N^{sl}_{vehicle}$ V the N^{sl}_{people} .
4	The N_{people} and the N_{people} V the N^{sl}_{people} .
5	The $N^{sl}_{quantity}$ of N^{pl}_{people} P the $N^{sl}_{vehicle}$ V the N^{sl}_{people} .
6	The N_{people} V that the N^{pl}_{people} V.
7	The N_{people} Adv V that the N^{pl}_{people} V .
8	The N_{people} V that the N^{pl}_{people} V Adv .
9	The N_{people} that V V the N^{sl}_{people} .
10	The N_{people} that V Pro V the N^{sl}_{people} .

(a) Synthetic templates

n	Template
1,2,3	The N_{people} $VP_{1,2,3}$. *The men are gon na have to move off-camera .*
4,5	The N_{people} read(s) an article about $NP_{1,2}$. *The man reads an article about the development of ascites in rats with liver cirrhosis .*
6,7	An article about $NP_{3,4}$ is read by N_{people} . *An article about the criterion on price stability , which was 27 % , is read by the child .*
8,9,10	Did the N_{people} hear about $NP_{5,6,7}$? *Did the teacher hear about the march on Employment which happened here on Sunday ?*

(b) Semi-natural templates

Table 1: The synthetic and semi-natural templates, with POS tags of the lexical items varied shown in blue with the plurality as superscript and the subcategory as subscript. The OPUS-extracted NP and VP fragments are red.

compositionality (Jacobson, 2002; Szabó, 2012). Under the local interpretation, natural language can hardly be considered compositional – many frequent phenomena such as homonyms, idioms and scope ambiguities cannot be resolved locally (Pagin and Westerståhl, 2010; Pavlick and Callison-Burch, 2016). The global interpretation handles such cases straightforwardly but does not match up with many a person's intuitions about the compositionality of language. After all, how useful is compositionality if composing the meanings of parts requires the entire rest of the sentence? This paradox inspired debates on the compositionality of natural language and is also highly relevant in the context of evaluating compositionality in neural models.

Previous compositionality tests (§6) considered only the local interpretation of compositionality, but to what extent is that relevant given the type of compositionality actually required to model natural language? Here, we aim to open up the discussion about what it means for computational models of language to be compositional by considering properties that require composing meaning locally as well as globally and evaluating them in models trained on unadapted natural language corpora.

3 Setup

3.1 Model and training

We focus on English-Dutch translation, for which we can ensure good command for both languages. We train Transformer-base models (Vaswani et al., 2017) using Fairseq (Ott et al., 2019). Our training data consists of a collection of MT corpora bundled in OPUS (Tiedemann and Thottingal, 2020), of which we use the English-Dutch subset provided by Tiedemann (2020), which contains 69M sentence pairs.[5] To examine the impact of the amount of training data – a dimension that is relevant because compositionality is hypothesised to be more important when resources are scarcer – we train one setup using the **full** dataset, one using $\frac{1}{8}$ of the data (**medium**), and one using one million source-target pairs in the **small** setup. For each setup, we train models with five seeds and average the results.

To evaluate our trained models, we adopt FLORES-101 (Goyal et al., 2021), which contains 3001 sentences from Wikinews, Wikijunior and WikiVoyage, translated by professional translators, split across three subsets. We train the models until convergence on the 'dev' set. Afterwards, we compute SacreBLEU scores on the 'devtest' set (Post, 2018), using beam search (beam size = 5), yielding scores of $20.6 \pm .4$, $24.4 \pm .3$ and $25.8 \pm .1$ for the small, medium and full datasets, respectively.[6]

3.2 Evaluation data

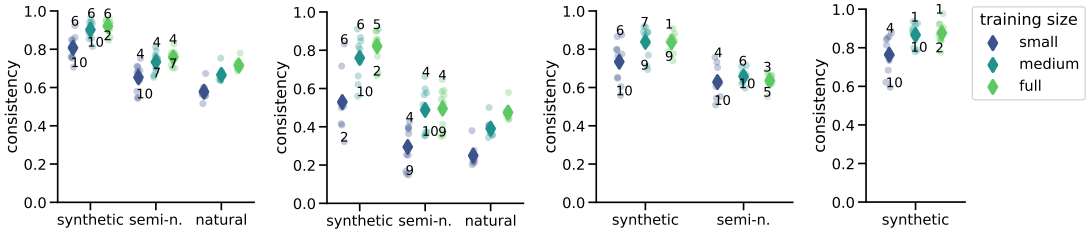
While all our models are trained on fully natural data, for evaluation we use different types of data: synthetic, semi-natural and natural data.

Synthetic data For our **synthetic** evaluation data, we consider the data generated by Lakretz et al. (2019), previously used to probe for hierarchical structure in neural language models. This data consist of sentences with a fixed syntactic structure and diverse lexical material. We extend the vocabulary and the templates used to generate the data and generate 3000 sentences for each of the resulting 10 templates (see Table 1a).

Semi-natural data In the synthetic data, we have full control over the sentence structure and lexical items, but the sentences are shorter (9 tokens vs 16 in OPUS) and simpler than typical in NMT data. To obtain more complex yet plausible test sentences, we employ a data-driven approach

[5] Visit the Tatoeba challenge for the OPUS training data.

[6] All training details are listed in Appendix E.



(a) $S_1 \rightarrow S'_1$ (b) $S_1 \rightarrow S_3$ (c) NP $\rightarrow$ NP′ (d) VP $\rightarrow$ VP′

Figure 1: Systematicity results for setup S → S CONJ S (a and b) and S → NP VP (c and d). Consistency scores are shown per evaluation data type (x-axis) and training dataset size (colours). Data points represent templates (○) and means over templates (◇).

S→S CONJ S

The girl sees that the men cry , and the poet criticises the king

The painter avoids the mayor , and the poet criticises the king

S→NP VP

NP→NP′	VP→VP′
The girl sees that the men cry	The girl sees that the men cry
The baker sees that the men cry	The girl sees that the aunts cry

Figure 2: Illustration of the systematicity experiments S → S CONJ S ($S_1 \rightarrow S_3$ is shown) and S → NP VP (both versions are shown). Each experiment involves extracting translations before and after the replacement of the blue part, and then comparing the translation of the underlined words.

to generate **semi-natural** data. Using the tree substitution grammar Double DOP (Van Cranenburgh et al., 2016), we obtain noun and verb phrases (NP, VP) whose structures frequently occur in OPUS. We then embed these NPs and VPs in ten synthetic templates with 3000 samples each (see Table 1b). See Appendix A for details on the data generation.

Natural data Lastly, we extract **natural** data directly from OPUS, as detailed in the subsections of the individual tests (§4).

4 Experiments and results

In our experiments, we consider *systematicity* (§4.1) and *substitutivity* (§4.2), to test for local compositionality, and *idiom translation* to probe for a more global type of processing (§4.3).

4.1 Systematicity

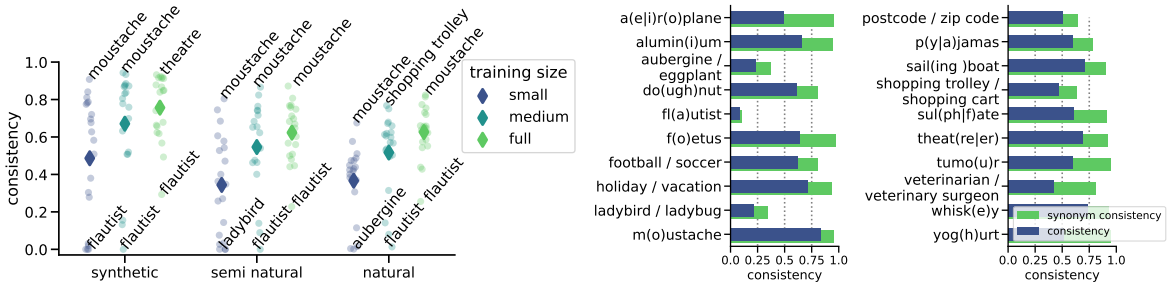
One of the most commonly tested properties of compositional generalisation is **systematicity** – the ability to understand novel combinations made up from known components (most famously, Lake and Baroni, 2018). In natural data, the number of potential recombinations to consider is infinite. We chose to focus on recombinations in two sentence-level context-free rules: S → NP VP and S → S CONJ S.

4.1.1 Experiments

Test design The first setup, S → NP VP, concerns recombinations of noun and verb phrases. We extract translations for input sentences from the templates from §3.2, as well as versions of them with the (1) noun (NP → NP') or (2) verb phrase (VP → VP') adapted. In (1), a noun from the NP in the subject position is replaced with a different noun while preserving number agreement with the VP. In (2), a noun in the VP is replaced. NP → NP' is applied to both synthetic and semi-natural data; VP → VP' only to synthetic data. We use 500 samples per template per condition per data type.

The second setup, S → S CONJ S, involves phrases concatenated using "and", and tests whether the translation of the second sentence is dependent on the first sentence. We concatenate two sentences (S_1 and S_2) from different templates, and we consider again two different conditions. First, in condition $S_1 \rightarrow S'_1$, we make a minimal change to S_1 yielding S'_1 by changing the noun in its verb phrase. In $S_1 \rightarrow S_3$, instead, we replace S_1 with a sentence S_3 that is sampled from a template different from S_1. We compare the translation of S_2 in all conditions. For consistency, the first conjunct is always sampled from the synthetic data templates. The second conjunct is sampled from synthetic data, semi-natural data, or from natural sentences sampled from OPUS with similar lengths and word-frequencies as the semi-natural inputs. We use 500 samples per template per condition per data type. Figure 2 provides an illustration of the different setups experimented with.

Evaluation In artificial domains, systematicity is evaluated by leaving out combinations of 'known components' from the training data and using them for testing purposes. The necessary familiarity of the components (the fact that they are 'known') is ensured by high training accuracies, and systematicity is quantified by measuring the test set accu-



(a) (b)

Figure 3: (a) Consistency scores of synonyms (averaged ◇, and per synonym ○) for substitutivity per evaluation data type, for three training set sizes. (b) Consistency per synonym, measured using full sentences (in dark blue) or the synonym's translation only (in green), averaged over training dataset sizes and data types.

racy. If the training data is a natural corpus and the model is evaluated with a measure like BLEU in MT, this strategy is not available. We observe that being systematic requires being consistent in the interpretation assigned to a (sub)expression across contexts, both in artificial and natural domains. Here, we, therefore, focus on **consistency** rather than accuracy, allowing us to employ a model-driven approach that evaluates the model's systematicity as the consistency of the translations when presenting words or phrases in multiple contexts.

We measure consistency as the equality of two translations after accounting for anticipated changes. For instance, in the S → NP VP setup, two translations are consistent if they differ in one word only, after accounting for determiner changes in Dutch ("de" vs "het"). In the evaluation of S → S CONJ S, we measure the consistency of the translations of the second conjunct.

4.1.2 Results

Figure 1 shows the results for the S → NP VP and S → S CONJ S setups (numbers available in Appendix B). The average performance for the natural data closely resembles the performance on *semi-natural* data, suggesting that the increased degree of control did not severely impact the results obtained using this generated data.[7] In general, the consistency scores are low, illustrating that models are prone to changing their translation of a (sub)sentence after small (unrelated) adaptations to the input. It hardly matters whether that change occurs in the sentence itself (S → NP VP), or in the other conjunct (S → S CONJ S), suggesting that the processing of the models is not local as assumed in strong compositionality. Models trained on more data seem more locally compositional, a somewhat contradictory solution to achieving compositionality, which, after all, is assumed to underlie the ability to generalise usage from *few* examples (Lake et al., 2019). This trend is also at odds with the hypothesis that inconsistencies are a consequence of the natural variation of language, which models trained on *more* data are expected to better capture.

[7] In our manual analysis (§5), however, we did observe a slightly different distribution of changes between these setups.

4.2 Substitutivity

Under a local interpretation of the principle of compositionality, synonym substitutions should be meaning-preserving: substituting a constituent in a complex expression with a synonym should not alter the complex expression's meaning, or, in the case of MT, its translation. Here, we test to what extent models' translations abide by this principle, by performing the **substitutivity** test from Hupkes et al. (2020), that measures whether the outputs remain consistent after synonym substitution.

4.2.1 Experiments

To find synonyms – source terms that translate into the same target terms – we exploit the fact that OPUS contains texts both in British and American English. Therefore, it contains synonymous terms that are spelt different – e.g. "doughnut" / "donut" – and synonymous terms with a very different form – e.g. "aubergine" / "eggplant". We use 20 synonym pairs in total (see Figure 3b).

Test design Per synonym pair, we select natural data from OPUS in which the terms appear and perform synonym substitutions. Thus, each sample has two sentences, one with the British and one with the American English term. We also insert the synonyms into the synthetic and semi-natural data using 500 samples per synonym pair per template, through subordinate clauses that modify a noun – e.g. "the king *that eats the doughnut*". In Appendix C, Table 6, we list all clauses used.

Evaluation Like systematicity, we evaluate substitutivity using the consistency score, expressing whether the model translations for a sample are identical. We report both the full sentence consistency and the consistency of the synonyms' translations only, excluding the context. Cases in which the model omits the synonym from both translations are labelled as consistent if the rest of the translation is the same for both input sequences.

4.2.2 Results

In Figure 3a, we summarise all substitutivity consistency scores (tables are in Appendix C). We observe trends similar to the systematicity results: models trained on larger training sets perform better and synthetic data yields more consistent translations compared to (semi-)natural data. We further observe large variations across synonyms, for which we further detail the performance aggregated across experimental setups in Figure 3b. The three lowest scoring synonyms – "flautist", "aubergine" and "ladybug" – are among the least frequent synonyms (see Appendix C), which stresses the importance of frequency for the model to pick up on synonymy.

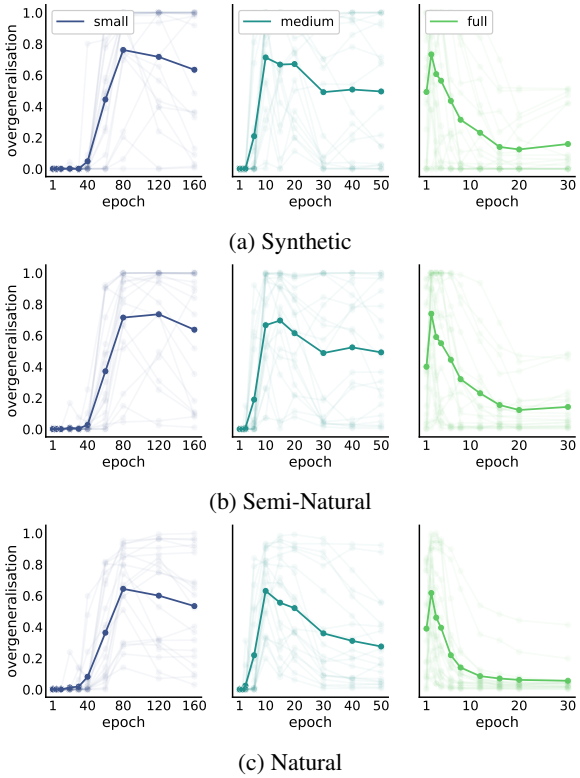
In Figure 3b, we show both the regular consistency and the consistency of the synonym translations, illustrating that a substantial part of the inconsistencies are due to varying translations of the context rather than the synonym itself, stressing again the non-local processing of the models.

4.3 Global compositionality

In our final test, we focus on exceptions to compositional rules. In natural language, typical exceptions that constitute a challenge for local compositionality are *idioms*. For instance, the idiom "raining cats and dogs" should be treated globally to arrive at its meaning of heavy rainfall. A local approach would yield an overly literal, non-sensical translation ("het regent katten en honden"). When a model's translation is too local, we follow Hupkes et al. (2020) in saying that it **overgeneralises**, or, in other words, it applies a general rule to an expression that is an exception to this rule. Overgeneralisation indicates that a language learner has internalised the general rule (e.g. Penke, 2012).

4.3.1 Experiments

We select 20 English idioms for which an accurate Dutch translation differs from the literal translation from the English MAGPIE corpus (Haagsma et al., 2020). Because acquisition of idioms is dependent on their frequency in the corpus, we use idioms



(a) Synthetic

(b) Semi-Natural

(c) Natural

Figure 4: Visualisation of overgeneralisation for idioms throughout training, with a line per idiom and the overall mean. Overgeneralisation occurs early on in training and precedes memorisation of idioms' translations. The colours indicate different training dataset sizes.

with at least 200 occurrences in OPUS based on exact matches, for which over 80% of the target translations does not contain a literal translation.

Test design Per idiom, we extract *natural* sentences containing the idiom from OPUS. For the synthetic and semi-natural data types, we insert the idiom in 500 samples per idiom per template, by attaching a subordinate clause to a noun – e.g. "the king *that said 'I knew the formula* ***by heart***'". The clauses used can be found in Appendix D, Table 7.

Evaluation Per idiom, we assess how often a model overgeneralises and how often it translates the idiom globally. To do so, we identify keywords that indicate that a translation is translated locally (literal) instead of globally (idiomatic). If the keywords' literal translations are present, the translation is labelled as an overgeneralised translation. For instance, for "by heart", the presence of "hart" ("heart") suggests a literal translation. An adequate paraphrase would say "uit het hoofd" ("from the head"). See Appendix D, Table 7, for the full list of keywords. We evaluate overgeneralisation for ten intermediate training checkpoints.

4.3.2 Results

In Figure 4, we report our results.[8] For all evaluation data types and all training set sizes, three phases can be identified. Initially, the translations do not contain the idiom's keyword, not because the idiom's meaning is paraphrased in the translation, but because the translations consist of high-frequency words in the target language only. Afterwards, overgeneralisation peaks: the model emits a very literal translation of the idiom. Finally, the model starts to memorise the idiom's translation. This is in accordance with results from Hupkes et al. (2020), and earlier results presented in the past tense debate by – among others – Rumelhart and McClelland (1986).

Although the height of the overgeneralisation peak is similar across evaluation data types and training set sizes, overgeneralisation is more prominent in converged models trained on smaller datasets than it is in models trained on the full corpus.[9] In addition to training dataset size, the type of evaluation data used also matters: there is more overgeneralisation for synthetic and semi-natural data compared to natural data, stressing the impact of the context in which an idiom is embedded. The extreme case of a context unsupportive of an idiomatic interpretation is a sequence of random words. To evaluate the hypothesis that this yields local translations, we surround the idioms with ten random words. The results (Appendix D, Table 7) indicate that, indeed, when the context provides no support at all for a global interpretation, the model provides a local translation for nearly all idioms. Overall, the results of this test provide an interesting contrast with our substitutivity and systematicity results: where in those tests, we saw processing that was *less local* than we expected, here, the behaviour shown by the models is instead *not global enough*.

5 Manual analysis

Our systematicity and substitutivity results demonstrate that models are not behaving compositional according to a strict definition of compositionality. However, we ourselves have argued that strict compositionality is not always appropriate to handle natural language. A reasonable question to ask is thus: are the inconsistencies we marked as non-compositional actually incorrect?

[8] Note that epochs consist of different numbers of samples: 1M, 8.6M and 69M for small, medium and full. Appendix D further details numerical results per idiom.

[9] Convergence is based on BLEU scores for validation data.

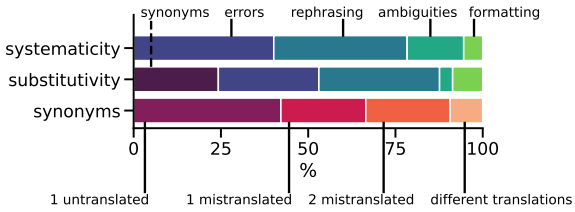


Figure 5: Relative frequencies of manually labelled inconsistencies in translations, averaged over data types and training set sizes. The 'synonyms' distribution further details the category 'synonyms' from row two.

Annotation setup To address this question, we perform a manual analysis. We annotate 900 inconsistent translation pairs of the systematicity and substitutivity tests to establish whether the inconsistencies are benign or concerning. We consider four different types of changes:

1. cases of *rephrasing*, where both translations are equally (in)correct;
2. changes reflecting different interpretations of *source ambiguities*;
3. cases in which one of the two translations contains an *error*;
4. *formatting* (mostly punctuation) changes.

For substitutivity samples, we also annotate whether the changes are related to the translation of the synonym, where we distinguish cases where

i. one of the synonym translations is incorrect;
ii. both are incorrect but in a different manner;
iii. both are correct but translated differently;
iv. one synonym remains untranslated.

We annotate all changes observed per pair and report the relative frequency per class. We summarise the results, aggregated over different training set sizes and the three data types, in Figure 5. For a more elaborate analysis and a breakdown per model and data type, we refer to Appendix F.

Results In the systematicity test, 40% of the marked inconsistencies reflects wrongfully translated parts in one of the two sentences, whereas 38% contains examples of rephrasing, 16% reflects ambiguities in the source sentences and 6% is caused by formatting differences. For substitutivity, most inconsistencies are similar to the ones observed in systematicity: only 24% involves the synonyms' translations, where one of them being untranslated was the most frequent category.

The distribution of these types of inconsistencies differ strongly per training data type. For models trained on less data, inconsistencies are more likely to represent errors, whereas models trained on more data rephrase more often. This result emphasises

that for lower-resource settings, being compositional is particularly relevant.

Another demonstration of this relevance comes from the observation that although models *can* emit correct translations for nearly all synonyms,[10] they do not always do so, depending on the context. To give a peculiar example: in "The child admires the king that eats the {doughnut, donut}", the snack was occasionally translated as "ezel" ("donkey").

Robustness and predictability Finally, we would like to stress that while rephrasing often might seem benign rather than concerning from the perspective of emitting adequate translations, its harmlessness still deserves some thought. There is a fine line between rephrasing and mistranslating: whether "the *single largest* business establishment" is referred to as "de grootste" ("the largest") or "de enige grootste" ("the only largest") may make or break a translation. Furthermore, if changes are unrelated to the contextual change (e.g. replacing "soccer" with "football"), this can be undesirable from a robustness and reliability perspective. This point becomes even more pronounced in cases where both translations are correct but have a different meaning. To analyse the extent to which inconsistencies are actually unmotivated, we investigated if we could trace them back to the contextual change, in particular focusing on whether changing synonyms from British to American spelling or vice versa might trigger a change in style or tone. We could not find evidence of such motivations, indicating that even correct cases of rephrasing were not caused by contextual changes that were *necessary* to take into account.

6 Related work

In previous work, a variety of artificial tasks have been proposed to evaluate compositional generalisation using non-i.i.d. test sets that are designed to assess a specific characteristic of compositional behaviour. Examples are *systematicity* (Lake and Baroni, 2018; Bastings et al., 2018; Hupkes et al., 2020), *substitutivity* (Mul and Zuidema, 2019; Hupkes et al., 2020), *localism* (Hupkes et al., 2020; Saphra and Lopez, 2020), *productivity* (Lake and Baroni, 2018) or *overgeneralisation* (Korrel et al., 2019; Hupkes et al., 2020; Dankers et al., 2021). Generally, neural models struggle to generalise in such evaluation setups.

There are also studies that consider compositional generalisation on more natural data. Such studies typically focus on either MT (Lake and Baroni, 2018; Raunak et al., 2019; Li et al., 2021) or semantic parsing (Finegan-Dollak et al., 2018; Keysers et al., 2019; Kim and Linzen, 2020; Shaw et al., 2021). Most of these studies consider small and highly controlled subsets of natural language.

Instead, we focus on models trained on fully natural MT datasets, which we believe to be the setup for compositionality evaluation that does most justice to the complexity of natural language: contrary to semantic parsing, where the outputs are structures created by expert annotators, in translation both inputs and outputs are fully-fledged natural language sentences. To the best of our knowledge, the only attempt to explicitly measure compositional generalisation of NMT models trained on large natural MT corpora is the study presented by Raunak et al. (2019). They measure productivity – generalisation to longer sentence lengths – of an LSTM-based NMT model trained on a full-size, natural MT dataset. Other studies using NMT, instead, consider toy datasets generated via templating (Lake and Baroni, 2018) or focus on short sentences excluding more complex constructions that contribute to the complexity of natural language for compositional generalisation, such as polysemous words or metaphors (Li et al., 2021).

7 Discussion

Whether neural networks can generalise compositionally is often studied using artificial tasks that assume strictly *local* interpretations of compositionality. We argued that such interpretations exclude large parts of language and that to move towards human-like productive usage of language, tests are needed that assess how compositional models trained on *natural data* are.[11] We laid out reformulations of three compositional generalisation tests – systematicity, substitutivity and overgeneralisation – for NMT models trained on natural corpora, and assessed models trained on different amounts of data. Our work provides an empirical contribution but also highlights vital hurdles to overcome when considering what it means for models of natural language to be compositional. Below, we reflect on these hurdles and our results.

[10] Apart from the model with the small training dataset that cannot translate "flautist" and "ladybug".

[11] Dupoux (2018) makes a similar point for models of language acquisition, providing several concrete examples where using less than fully complex data proved problematic.

The proxy-to-meaning problem Compositionality is a property of the mapping between the form and meaning of an expression. Since translation is a *meaning-preserving* mapping from form in one language to form in another, it is an attractive task to evaluate compositionality: the translation of its sentence can be seen as a proxy to its meaning. However, while expressions are assumed to have only one meaning, translation is a *many-to-many* mapping: the same sentence can have multiple correct translations. This does not only complicate evaluation – MT systems are typically evaluated with BLEU because accuracy is not a suitable option – it also raises questions about how compositional the desired behaviour of an MT model should be. On the one hand, one could argue that for optimal generalisation, robustness, and accountability, we like models to behave systematically and consistently: we expect the translations of expressions to be independent of unrelated contextual changes that do not affect their meaning (e.g. swapping out a synonym in a nearby sentence). Additionally, model performance could be improved if small changes do not introduce errors in unrelated parts of the translation. On the other hand, non-compositional behaviour is not always incorrect – it is one of the main arguments in our plea to test compositionality 'in the wild' – and we observe that indeed, not all non-compositional changes alter the correctness of the resulting translations. Changing a translation from "atleet" ("athlete") to "sporter" ("sportsman") based on an unrelated word somewhat far away may not be (locally) compositional, but is it a problem? And how do we separate such 'harmful' mistakes from helpful ones?

The locality problem Inextricably linked to the proxy-to-meaning problem is the locality problem. In our tests we see that *small, local source changes* elicit *global changes in translations*. For instance, in our systematicity tests, changing one noun in a sentence elicited changes in the translation of a sentence that it was conjoined with. In our substitutivity test, even synonyms that merely differed in spelling (e.g. "doughnut" and "donut") elicited changes to the remainder of the sentence. This counters the idea of compositionality as a means of productively reusing language: if a phrase's translation depends on (unrelated) context that is not in its direct vicinity, this suggests that more evidence is required to acquire the translation of this phrase.

Tests involving synthetic data present the models with sentences in which maximally local behaviour is possible, and we argue that it is, therefore, also desirable. Our experiments show that even in such setups, models do not translate in a local fashion: with varying degrees of correctness, they frequently change their translation when we slightly adapt the input. On the one hand, this well-known *volatility* (see also Fadaee and Monz, 2020) might be essential for coping with ambiguities for which meanings are context-dependent. On the other hand, our manual analysis shows that the observed non-compositional behaviour does not reflect the incorporation of necessary contextual information and that oftentimes it is even altering the correctness of the translations. Furthermore, this erratic behaviour highlights a lack of default reasoning, which can, in some cases, be problematic or even harmful, especially if faithfulness (Parthasarathi et al., 2021) or consistency is important.

In linguistics, it has been discussed how to extend the syntax and semantics such that 'problem cases' can be a part of a compositional language (Westerståhl, 2002; Pagin and Westerståhl, 2010). In such formalisations, global information is used to disambiguate the problem cases, while other parts of the language are still treated locally. In our models, global behaviour appears in situations where a local treatment would be perfectly suitable and where there is no clear evidence for ambiguity. We follow Baggio (2021) in suggesting that we should learn from strategies employed by humans, who can assign compositional interpretations to expressions but can for some inputs also derive non-compositional meanings. For *human-like* linguistic generalisation, it is vital to investigate how models can represent both these types of processing, providing a locally compositional treatment when possible and deviating from that when needed.

Conclusion In conclusion, with this work, we contribute to the question of how compositional models trained on *natural* data are, and we argue that MT is a suitable and relevant testing ground to ask this question. Focusing on the balance between *local* and *global* forms of compositionality, we formulate three different compositionality tests and discuss the issues and considerations that come up when considering compositionality in the context of natural data. Our tests indicate that models show both local and global processing, but not necessarily for the right samples. Furthermore, they underscore the difficulty of separating helpful and harmful types of non-compositionality, stressing the need to rethink the evaluation of compositionality using natural language, where composing meaning is not as straightforward as doing the math.

Acknowledgements

We thank Sebastian Riedel, Douwe Kiela, Thomas Wolf, Khalil Sima'an, Marzieh Fadaee, Marco Baroni, Brenden Lake and Adina Williams for providing feedback on this draft and our work in several different stages of it. We thank Michiel van der Meer for contributing to the initial experiments that led to this paper. A special thanks goes to Angela Fan, who assisted us at several points to get the ins and outs of training large MT models and double-checked several steps of our pipeline and to our ARR reviewers, who provided amazingly high quality feedback. 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.

References

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

Jasmijn Bastings, Marco Baroni, Jason Weston, Kyunghyun Cho, and Douwe Kiela. 2018. Jump to better conclusions: SCAN both left and right. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 47–55. Association for Computational Linguistics.

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.

Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to German plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108, Online. Association for Computational Linguistics.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Marzieh Fadaee and Christof Monz. 2020. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation, NGT@ACL 2020, Online, July 5-10, 2020*, pages 88–96. Association for Computational Linguistics.

Catherine Finegan-Dollak, Jonathan K Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360.

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

Eduardo García-Ramírez. 2019. *Open Compositionality: Toward a New Methodology of Language*. Rowman & Littlefield.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

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.

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.

Pauline Jacobson. 2002. The (dis)organization of the grammar: 25 years. *Linguistics and Philosophy*, 25(5/6):601–626.

Theo MV Janssen. 1998. Algebraic translations, correctness and algebraic compiler construction. *Theoretical Computer Science*, 199(1-2):25–56.

Theo MV Janssen and Barbara H Partee. 1997. Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *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.

Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. 2019. Transcoding compositionally: Using attention to find more generalizable solutions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 1–11.

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.

Brenden M. Lake, Tal Linzen, and Marco Baroni. 2019. Human few-shot learning of compositional instructions. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, pages 611–617. cognitivesciencesociety.org.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20.

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.

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

Mathijs Mul and Willem Zuidema. 2019. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization. In *CoRR, abs/1906.00180*.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 1–9. Association for Computational Linguistics.

Peter Pagin and Dag Westerståhl. 2010. Compositionality ii: Arguments and problems. *Philosophy Compass*, 5(3):265–282.

Barbara Partee. 1984. Compositionality. *Varieties of formal semantics*, 3:281–311.

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.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173.

Martina Penke. 2012. The dual-mechanism debate. In *The Oxford handbook of compositionality*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

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*.

MT Rosetta. 1994. The rosetta characteristics. In *Compositional Translation*, pages 85–102. Springer.

D E Rumelhart and J McClelland. 1986. On Learning the Past Tenses of English Verbs. In *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 216–271. MIT Press, Cambridge, MA.

Naomi Saphra and Adam Lopez. 2020. LSTMs compose—and Learn—Bottom-up. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2797–2809, Online. Association for Computational Linguistics.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? 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 922–938, Online. Association for Computational Linguistics.

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

Zoltan Szabó. 2012. The case for compositionality. *The Oxford handbook of compositionality*, 64:80.

Jörg Tiedemann. 2020. The Tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

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.

Andreas Van Cranenburgh, Remko Scha, and Rens Bod. 2016. Data-oriented parsing with discontinuous constituents and function tags. *Journal of Language Modelling*, 4(1):57–111.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Dag Westerståhl. 2002. On the compositionality of idioms. *Proceedings of LLC8. CSLI Publications*.

Appendix A Semi-natural templates

The semi-natural data that we use in our test sets is generated with the library `DiscoDOP`,[12] developed for data-oriented parsing (Van Cranenburgh et al., 2016). We generate the data with the following seven step process:

Step 1. Sample 100k English OPUS sentences.

Step 2. Generate a treebank using the disco-dop library and the `discodop parser en_ptb` command. The library was developed for discontinuous data-oriented parsing. Use the library's `--fmt bracket` to turn off discontinuous parsing.

Step 3. Compute tree fragments from the resulting treebank (`discodop fragments`). These tree fragments are the building blocks of a Tree-Substitution Grammar.

Step 4. We assume the most frequent fragments to be common syntactic structures in English. To construct complex test sentences, we collect the 100 most frequent fragments containing at least 15 non-terminal nodes for NPs and VPs.

Step 5. Selection of three VP and five NP fragments to be used in our final semi-natural templates. These structures are selected through qualitative analysis for their diversity.

Step 6. Extract sentences matching the eight fragments (`discodop treesearch`).

Step 7. Create semi-natural sentences by varying one lexical item and varying the matching NPs and VPs retrieved in Step 6.

In Table 2, we provide examples for each of the ten templates used, along with the internal structure of the complex NP or VP that is varied in the template. In Table 3, we provide some additional examples for our ten synthetic templates.

n	**Template**
1	The N_{people} (VP (TO) (VP (VB) (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP)))))))
	E.g. The woman wants to use the Internet as a means of communication .
2	The N_{people} (VP (VBP) (VP (VBG) (S (VP (TO) (VP (VB) (S (VP (TO) (VP))))))))
	E.g. The men are gon na have to move off-camera .
3	The N_{people} (VP (VB) (NP (NP) (PP (IN) (NP))) (PP (IN) (NP (NP) (PP (IN) (NP)))))
	E.g. The doctors retain 10 % of these amounts by way of collection costs .
4	The N_{people} reads an article about (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP)))))))
	E.g. The friend reads an article about the development of ascites in rats with liver cirrhosis .
5	The N_{people} reads an article about (NP (NP (DT) (NN)) (PP (IN) (NP (NP) (SBAR (S (WHNP (WDT)) (VP)))))) .
	E.g. The teachers read an article about the degree of progress that can be achieved by the industry .
6	An article about (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP))))))) is read by the N_{people} .
	E.g. An article about the inland transport of dangerous goods from a variety of Member States is read by the lawyer .
7	An article about (NP (NP) (PP (IN) (NP (NP) (, ,) (SBAR (S (WHNP (WDT)) (VP)))))) , is read by the N_{people} .
	E.g. An article about the criterion on price stability , which was 27 % , is read by the child .
8	Did the N_{people} hear about (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP (NP) (PP (IN) (NP))))))) .
	E.g. Did the friend hear about an inhospitable fringe of land on the shores of the Dead Sea ?
9	Did the N_{people} hear about (NP (NP (DT) (NN)) (PP (IN) (NP (NP) (SBAR (S (WHNP (WDT)) (VP)))))) ?
	E.g. Did the teacher hear about the march on Employment which happened here on Sunday ?
10	Did the N_{people} hear about (NP (NP) (SBAR (S (VP (TO) (VP (VB) (NP (NP) (PP (IN) (NP))))))))) ?
	E.g. Did the lawyers hear about a qualification procedure to examine the suitability of the applicants ?

Table 2: Semi-natural data templates along with their identifiers (n). The syntactic structures for noun and verb phrases in purple are instantiated with data from the OPUS collection. Generated data from every template contains varying sentence structures and varying tokens but the predefined tokens in black remain the same.

[12] https://github.com/andreasvc/disco-dop

n	**Template**
1	The $\mathrm{N}_{\mathrm{people}}$ $\mathrm{V}_{\mathrm{transitive}}$ the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{people}}$. *E.g. The poet criticises the king .*
2	The $\mathrm{N}_{\mathrm{people}}$ Adv $\mathrm{V}_{\mathrm{transitive}}$ the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{people}}$. *E.g. The victim carefully observes the queen .*
3	The $\mathrm{N}_{\mathrm{people}}$ P the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{vehicle}}$ $\mathrm{V}_{\mathrm{transitive}}$ the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{people}}$. *E.g. The athlete near the bike observes the leader .*
4	The $\mathrm{N}_{\mathrm{people}}$ and the $\mathrm{N}_{\mathrm{people}}$ $\mathrm{V}^{\mathrm{pl}}_{\mathrm{transitive}}$ the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{people}}$. *E.g. The poet and the child understand the mayor .*
5	The $\mathrm{N}^{\mathrm{sl}}_{\mathrm{quantity}}$ of $\mathrm{N}^{\mathrm{pl}}_{\mathrm{people}}$ P the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{vehicle}}$ $\mathrm{V}^{\mathrm{sl}}_{\mathrm{transitive}}$ the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{people}}$. *E.g. The group of friends beside the bike forgets the queen .*
6	The $\mathrm{N}_{\mathrm{people}}$ $\mathrm{V}_{\mathrm{transitive}}$ that the $\mathrm{N}^{\mathrm{pl}}_{\mathrm{people}}$ $\mathrm{V}^{\mathrm{pl}}_{\mathrm{intransitive}}$. *E.g. The farmer sees that the lawyers cry .*
7	The $\mathrm{N}_{\mathrm{people}}$ Adv $\mathrm{V}_{\mathrm{transitive}}$ that the $\mathrm{N}^{\mathrm{pl}}_{\mathrm{people}}$ $\mathrm{V}^{\mathrm{pl}}_{\mathrm{intransitive}}$. *E.g. The mother probably thinks that the fathers scream .*
8	The $\mathrm{N}_{\mathrm{people}}$ $\mathrm{V}_{\mathrm{transitive}}$ that the $\mathrm{N}^{\mathrm{pl}}_{\mathrm{people}}$ $\mathrm{V}^{\mathrm{pl}}_{\mathrm{intransitive}}$ Adv . *E.g. The mother thinks that the fathers scream carefully .*
9	The $\mathrm{N}_{\mathrm{people}}$ that $\mathrm{V}_{\mathrm{intransitive}}$ $\mathrm{V}_{\mathrm{transitive}}$ the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{people}}$. *E.g. The poets that sleep understand the queen .*
10	The $\mathrm{N}_{\mathrm{people}}$ that $\mathrm{V}_{\mathrm{transitive}}$ Pro $\mathrm{V}^{\mathrm{sl}}_{\mathrm{transitive}}$ the $\mathrm{N}^{\mathrm{sl}}_{\mathrm{people}}$. *E.g. The mother that criticises him recognises the queen .*

Table 3: Synthetic sentence templates similar to Lakretz et al. (2019), along with their identifiers (n).

Appendix B Systematicity

Table 4 provides the numerical counterparts of the results visualised in Figure 1.

Data	**Condition**	**Model**		
		small	medium	full
S → NP VP				
synthetic	NP	.73	.84	.84
synthetic	VP	.76	.87	.88
semi-natural	NP	.63	.66	.64
S → S CONJ S				
synthetic	S'_1	.81	.90	.92
synthetic	S_3	.53	.76	.82
semi-natural	S'_1	.65	.73	.76
semi-natural	S_3	.29	.49	.49
natural	S'_1	.58	.67	.72
natural	S_3	.25	.39	.47

(a) Per models' training set size

Template									
1	2	3	4	5	6	7	8	9	10
.86	.74	.85	.87	.75	.89	.85	.85	.70	.68
.92	.73	.90	.91	.84	.88	.85	.82	.77	.74
.66	.63	.65	.70	.64	.69	.63	.63	.60	.58
.91	.82	.88	.88	.86	.95	.90	.91	.84	.79
.75	.54	.72	.66	.73	.88	.74	.81	.66	.55
.73	.75	.75	.80	.75	.73	.66	.68	.64	.64
.50	.50	.51	.58	.52	.43	.35	.31	.28	.29
.67	.74	.65	.64	.63	.64	.62	.66	.63	.66
.39	.49	.35	.35	.34	.37	.33	.38	.34	.38

(b) Per template

Table 4: Consistency scores for the systematicity experiments, detailed per experimental setup and evaluation data type. We provide scores (a) per models' training set size, and (b) per template of our generated evaluation data. For natural data, the template number is meaningless, apart from the fact that it determines sentence length and word frequency.

Appendix C Substitutivity

Synonyms employed In Table 5, we provide some information about the synonymous word pairs used in the substitutivity test, including their frequency in OPUS and their most common Dutch translation. The last column of the table contains the subordinate clauses that we used to include the synonyms in the synthetic and semi-natural data. We include them as a relative clause behind nouns representing a human, such as "The poet criticises the king that eats the doughnut".

Detecting synonym translations To find the span of text in the translation which is the translation of the synonym, we apply a relatively simple heuristic. We generate a number of short sentences such as "This is the NOUN", feed those to all our trained models, and extract the top-5 answers in the beam. We then use the list of all words resulting from this protocol – which we manually checked – to find synonym translations in the model output.

Results In the main paper, Figures 3a and 3b provided the consistency scores for the substitutivity tests. Here, Table 6 further details the results from the figure, by presenting the average consistency per evaluation data type and training set size, and per evaluation data type and synonym pair.

Synonym pair *British*	*Freq.*	*American*	*Freq.*	**Dutch translation**	**Subordinate clause**
aeroplane	6728	airplane	5403	vliegtuig	that travels by ...
aluminium	17982	aluminum	5700	aluminium	that sells ...
doughnut	2014	donut	1889	donut	that eats the ...
foetus	1943	fetus	1878	foetus	that researches the ...
flautist	112	flutist	101	fluitist	that knows the ...
moustache	1132	mustache	1639	snor	that has a ...
tumour	7338	tumor	6348	tumor	that has a ...
pyjamas	808	pajamas	1106	pyjama	that wears ...
sulphate	3776	sulfate	1143	zwavel	that sells ...
yoghurt	1467	yogurt	2070	yoghurt	that eats the ...
aubergine	765	eggplant	762	aubergine	that eats the ...
shopping trolley	217	shopping cart	13366	winkelwagen	that uses a ...
veterinary surgeon	941	veterinarian	6995	dierenarts	that knows the ...
sailing boat	5097	sailboat	1977	zeilboot	that owns a ...
football	33125	soccer	6841	voetbal	that plays ...
holiday	125430	vacation	23532	vakantie	that enjoys the ...
ladybird	235	ladybug	303	lieveheersbeestje	that caught a ...
theatre	19451	theater	13508	theater	that loves ...
postcode	479	zip code	1392	postcode	with the same ...
whisky	3604	whiskey	4313	whisky	that drinks ...

Table 5: Synonyms for the substitutivity test, along with their OPUS frequency, Dutch translation, and the subordinate clause used to insert them in the data.

Data	**Metric**	**Model** small	medium	full
synthetic	con.	.49	.67	.76
	syn. con.	.67	.82	.93
semi-natural	con.	.34	.55	.62
	syn. con.	.62	.84	.93
natural	con.	.37	.52	.63
	syn. con.	.61	.75	.85

(a) Per models' training set size

Data	**Metric**	**Synonym** aeroplane	aluminium	doughnut	foetus	flautist	moustache	tumour	pyjamas	sulphate	yoghurt	aubergine	shopping trolley	veterinary surgeon	sailing boat	football	holiday	ladybird	theatre	postcode	whisky
synthetic	con.	.54	.87	.74	.82	.10	.92	.78	.64	.79	.55	.25	.40	.64	.73	.68	.81	.27	.85	.48	.88
	syn. con.	1.0	1.0	.87	1.0	.10	1.0	1.0	.80	.95	1.0	.38	.48	.90	1.0	.75	1.0	.40	.99	.53	1.0
semi-natural	con.	.43	.59	.58	.54	.08	.85	.52	.55	.56	.42	.24	.31	.33	.73	.66	.71	.20	.62	.43	.75
	syn. con.	.99	.99	.83	1.0	.09	1.0	.98	.72	.90	.98	.40	.50	.77	1.0	.90	1.0	.38	.95	.58	.99
natural	con.	.50	.52	.53	.56	.09	.75	.50	.60	.47	.57	.23	.70	.29	.64	.55	.62	.17	.59	.61	.58
	syn. con.	.89	.85	.73	.91	.11	.87	.87	.82	.88	.86	.32	.92	.75	.71	.79	.81	.27	.82	.81	.80

(b) Per synonym

Table 6: Consistency scores for the substitutivity experiments, detailed per evaluation data type. We present scores (a) per models' training set size and (b) per synonym.

Appendix D Global compositionality

Idioms employed Table 7 provides more information on the idioms used in our global compositionality test. In the first column, we list all idioms we used, along with the *keywords* that we used to determine if their translation is local or not. To extract the natural data, we retrieved exact matches with OPUS source sentences. The idioms' keywords are mostly nouns that either translate into a different word in an accurate paraphrased translation in Dutch (e.g. "across the **board**" would be "over de hele linie"), or should disappear in the translation (e.g. "do the right **thing**" typically translates into "het juiste doen" in the corpus).

In the second column of Table 7, we list the subordinate clauses that we used to include idioms in the synthetic and semi-natural data. The clauses themselves are drawn from source sentences in OPUS. To incorporate them in synthetic and semi-natural sentences, we include them as a relative clause behind nouns representing a human, by attaching *"that said '[idiom]"'*. For instance: "The poet criticises the king that said 'Have you gone out of your mind'."

In the third column of Table 7, we show local translations of the idioms, elicited from the model by embedding the idiom in a string of ten random nouns. Even "out of the blue", which is rarely overgeneralised when presented in synthetic, semi-natural or natural contexts, is locally translated. This indicates that the idiom is not stored as one lexical unit per se but that it is only translated globally in specific contexts.

Results In the main paper, in Figure 4, we visualised how overgeneralisation changes over the course of training, averaged over idioms. In Table 8, we detail the maximum overgeneralisation observed per idiom.

Idiom	Subordinate clause	Local translation
once in a while	that said " I will play it once in a while "	eens in een tijdje
do the right thing	that said " Just do the right thing "	doen het juiste ding
out of your mind	that said " Have you gone out of your mind "	uit je hoofd
state of the art	that said " This is a state of the art, official facility "	stand van de kunst
from scratch	that said " We are cooking from scratch every day "	van kras
take stock	that said " Take stock of the lessons to be drawn "	nemen voorraad
across the board	that said " I got red lights all across the board "	aan boord
in the final analysis	that said " In the final analysis, this is what matters "	in de laatste analyse
out of the blue	that said " It just came out of the blue "	uit het blauwe
in tandem	that said " We will work with them in tandem "	in tandem
by heart	that said " I knew the formula by heart "	door hart
come to terms with	that said " I have come to terms with my evil past "	komen overeen met
by the same token	that said " By the same token I will oppose what is evil "	bij dezelfde token
at your fingertips	that said " The answer is right at your fingertips "	binnen handbereik
look the other way	that said " We cannot look the other way either "	kijken de andere manier
follow suit	that said " And many others follow suit "	volgen pak
keep tabs on	that said " I keep tabs on you "	houden tabs
in the short run	that said " In the short run it clearly must be "	in de korte lopen
by dint of	that said " We are part of it by dint of our commitment "	door de int
set eyes on	that said " I wish I had never set eyes on him "	set ogen op

Table 7: Idioms used in the overgeneralisation test. The words that are indicative of a local translation are underlined, we check for their presence to label a translation as an overgeneralisation. The listed subordinate clauses are used to insert the idioms into synthetic and semi-natural templates. The local translation indicated is the translation given by the model when the idiom is embedded in a string of ten random words.

Appendix E Reproducibility details

E.1 Data

Training data Our training data consists of the English-Dutch subset of the MT corpus OPUS (Tiedemann and Thottingal, 2020), provided by Tiedemann (2020). This data contains in total 69M source-target pairs. The data can be found on https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/data/README-v2020-07-28.md.

Data	Model	Idiom																			
		once in a while	do the right thing	out of your mind	state of the art	from scratch	take stock	across the board	in the final analysis	out of the blue	in tandem	by heart	come to terms with	by the same token	look the other way	at your fingertips	follow suit	keep tabs on	in the short run	by dint of	set eyes on
synthetic	small	.98	.92	.98	1.0	.40	.75	1.0	1.0	1.0	1.0	1.0	.01	1.0	1.0	1.0	.99	1.0	.72	.20	.74
	medium	.99	.96	.98	1.0	.76	.73	1.0	1.0	1.0	1.0	1.0	.22	1.0	1.0	1.0	1.0	.57	.55	.38	.57
	full	.97	.86	.97	1.0	.50	.56	1.0	1.0	1.0	1.0	1.0	.24	1.0	.91	1.0	1.0	.74	.38	.24	.44
semi-natural	small	.95	.66	.98	1.0	.49	.73	1.0	1.0	1.0	.97	1.0	.08	1.0	.98	1.0	.88	.99	.56	.15	.81
	medium	.91	.60	.95	1.0	.78	.63	.96	1.0	1.0	.97	1.0	.31	.99	.99	1.0	.97	.74	.45	.30	.59
	full	.97	.55	.95	1.0	.40	.68	.99	1.0	1.0	.99	1.0	.31	1.0	.90	1.0	.97	.90	.25	.23	.47
natural	small	.80	.51	.80	.97	.84	.31	.75	.96	.92	.82	.88	.14	.74	.60	1.0	.40	.96	.29	.23	.87
	medium	.80	.50	.82	.96	.84	.32	.71	.94	.92	.68	.90	.22	.74	.63	.99	.39	.61	.33	.29	.84
	full	.79	.39	.83	.95	.90	.36	.83	.98	.95	.89	.90	.11	.65	.55	1.0	.65	.56	.19	.27	.76

Table 8: Maximum overgeneralisation observed over the course of training, per evaluation data type, training set size and idiom.

Preprocessing We tokenise the data using the tokenisation script[13] from the SMT library Moses.[14] Following the number of subwords suggested by Tiedemann (2020), we generate a subword vocabulary applying 60k BPE merge-operations. To do so, we use the `learn_bpe.py` script provided in the SUBWORD_NMT[15] repository hosted by Rico Sennrich.

Different corpora We train models on three different sizes of corpora: SMALL, MEDIUM and FULL. To generate these corpora, we first shuffle the OPUS training data using the bash function `shuffle`. To generate the SMALL and MEDIUM corpora, we take the first 8582811 and 1072851 sentences of this shuffled corpus, which corresponds to $\frac{1}{8}$th and $\frac{1}{64}$th of the full training corpus, respectively. For each setting, we train models with seeds $\{1, 2, 3, 4, 5\}$.

Test and validation data Initially, we aimed to evaluate our models using the commonly used MT test sets OPUS-100[16] and the test partition of the TED talk corpus.[17] However, it turned out that both these test sets were almost fully contained in our training corpus. We, therefore, adopted the newer FLORES-101 corpus (Goyal et al., 2021), of which we used both the 'dev' and the 'devtest' set. The data can be downloaded from `https://dl.fbaipublicfiles.com/flores101/dataset/flores101_dataset.tar.gz`. To compute BLEU scores, we tokenised the data with the Moses tokenisation script mentioned above, and then used the commandline script `fairseq-generate` to compute scores.

We furthermore use several evaluation sets to assess the compositional abilities of our trained models. The data for these tests, as well as scripts to run them and plot their results, can be found in the following repository: `https://github.com/i-machine-think/compositionality_paradox_mt`.

E.2 Architecture and training

As reported in the main text, we focus on English-Dutch translation, and all our models are Transformer-base models, as implemented in Fairseq (Ott et al., 2019).[18] Both the encoder and the decoder of this model have an embedding dimension of 512, 6 layers, 8 attention heads and a feed-forward layer dimension of 2048. With our vocabulary, the models have a total of around 80M trainable parameters.

To train our models, we follow the training procedure suggested by Ott et al. (2018), which can be found at `https://github.com/pytorch/fairseq/tree/master/examples/scaling_nmt`. To summarise, we share all embeddings between the encoder and the decoder, use Adam as optimiser with β-values (0.9, 0.98),

[13] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl`

[14] `https://github.com/moses-smt/mosesdecoder`

[15] `https://github.com/rsennrich/subword-nmt/blob/master/subword_nmt/learn_bpe.py`

[16] `http://data.statmt.org/opus-100-corpus/v1.0/supervised/en-nl/`

[17] `https://github.com/neulab/word-embeddings-for-nmt`

[18] We used the implementation as it was on May 12, 2021: `https://github.com/pytorch/fairseq/blob/d151f2787240cca4e3c7e47640e647f8ae028c37/fairseq/models/transformer.py`

Training set size	Seed	BLEU dev	BLEU devtest
small	1	20.92	21.14
	2	20.77	20.37
	3	20.42	20.11
	4	20.95	20.23
	5	20.88	20.84
medium	1	24.09	24.18
	2	25.05	24.71
	3	24.55	24.42
	4	24.09	23.93
	5	24.55	24.10
full	1	26.17	25.63
	2	25.71	25.63
	3	25.82	25.72
	4	26.19	25.84
	5	25.86	25.76

Table 9: BLEU scores for the 'dev' and 'devtest' subsets of the FLORES datasets, for models trained on corpora of three sizes, for five seeds per training set size.

starting from an initial warmup learning rate of 1e-07 for 4000 warmup updates and a learning rate of 0.0005 afterwards, using inverse square root as the learning rate scheduler. We use a clip-norm of 0.0, dropout of 0.3, weight-decay of 0.0001, label-smoothing of 0.1. The maximum number of tokens in a batch is 3584, we simulate larger batches by increasing the update frequency to 8. To determine early stopping, we use a patience of 10 (i.e. we stop training if a model does not improve on the dev set anymore for 10 epochs, and take the best checkpoint at that point). Any other hyperparameters involved follow the Fairseq default. We provide the BLEU scores per model seed in Table 9.

E.3 Compute

All experiments were ran using Tesla V100 GPUs on an internal SLURM-based cluster. Training a transformer-base model on our small, medium and full dataset takes on average 3.5, 17 and 113 minutes per epoch, respectively (numbers are rounded) on 32 GPUs. This makes the total training time for these models, which are trained for around 160, 60 and 30 epochs, 10, 17 and 56 hours, respectively (again, spread over 32 GPUs).

Appendix F Manual analysis

Our quantitative tests provide information on when a model behaves locally and when globally in automated form but they do not consider whether that behaviour is incorrect or not. More simply put, we do not know whether the changes that we observe are actually resulting in incorrect translations. We complement these scores with an elaborate manual analysis, which provides more insight into the nature of the non-compositional behaviour we registered.

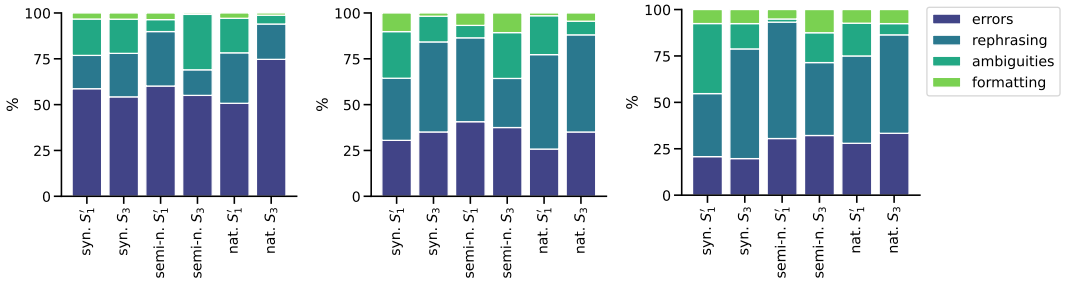
F.1 Setup

Data sampling We randomly sample 900 examples for substitutivity (100 for each {model}×{test data type} tuple) and 900 examples for systematicity (50 for each {model}×{test data type}×{S'_1, S_3} tuple), randomly distributed over templates. In all cases, we sample sentences randomly from the five seeds that we trained, and from all templates. For substitutivity, we sample five examples for each synonym for every {model}×{test data type} pair.

Annotation procedure For each of these samples, we annotate how they differ, where we distinguish between four general categories:

i. *Rephrasing*: part of the sentence is rephrased (but both phrases are equally (in)correct);
ii. *Source ambiguities*: there is an ambiguity in the source sentence, and the model switches its interpretation;
iii. *Errors*: one of the translations contains an error that the other one does not;
iv. *Formatting*: minor formatting changes, consisting mostly of insertions/deletions of punctuation.

For the substitutivity data, we separately annotate changes that are related to the translation of the synonym, where we distinguish cases in which both synonyms are correctly or incorrectly translated from cases in which one of the translations is correct. We annotate all changes observed in a sample – one sentence may thus contain annotations for multiple changes – and report the relative frequency of each class of errors.

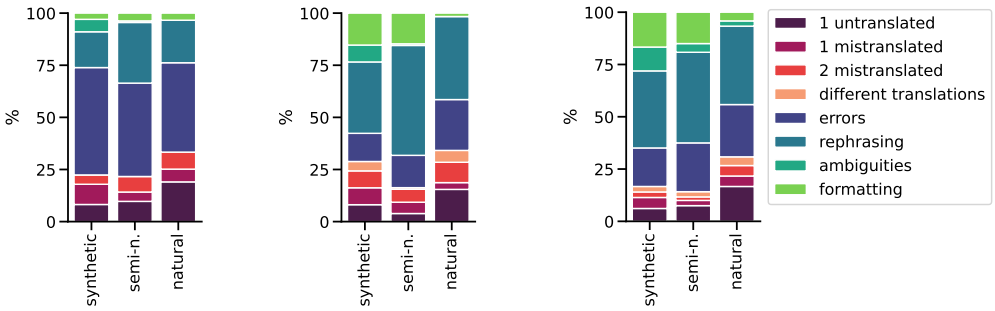


(a) Small training set (b) Medium training set (c) Full training set

Figure 6: Distribution of error types for sentences that contain inconsistencies in systematicity, detailed per model trained on the training set sizes in the subcaptions.

F.2 Results

We provide a summary of the results in Figure 6 for systematicity and Figure 7 for substitutivity. As a general trend, the results reflect that in models trained on smaller datasets, more mistakes are actually errors, rather than multiple correct alternatives. In the systematicity test, 59% of the inconsistencies for the models trained on the smallest dataset are erroneous changes, versus 34% and 27% in the models trained on the medium and largest dataset, when we average the percentages over the different subsets annotated. For substitutivity, the percentage of erroneous changes unrelated to the synonyms comprises 46%, 18% and 22% for the smallest, medium and full dataset, respectively. On top of that, there were inconsistencies related to the synonyms, that represented 26%, 26% and 21% for the three dataset sizes, respectively. While this is expected, to some extent, it still constitutes a problem: for models trained on smaller amounts of data, being able to translate in a compositional manner is particularly relevant. Below, we further elaborate on the types of inconsistencies encountered per annotation category, including some examples.



(a) Small training set (b) Medium training set (c) Full training set

Figure 7: Distribution of the types of inconsistencies observed in the substitutivity test, detailed per model trained on the training set sizes in the subcaptions. The red colour scheme represents error types specific to this experiment.

F.2.1 Rephrasing

A large portion of the inconsistencies concerns pairs where one translation can be considered a rephrased version of the other translation. A common cause of this is a **reordering of words** that does not impact the grammaticality or meaning of the Dutch sentence – e.g. in sentences with adverbs ("heeft de burgemeester zeker in de gaten" vs "heeft zeker de burgemeester in de gaten") or relative clauses with direct objects ("die genieten van de vakantie" vs "die van de vakantie genieten"). We could not trace these reorderings back to the specific change made in the systematicity or substitutivity tests. Consider, for instance, Example (1), where the reordering happens as a consequence of changing the word "king" to "father". Note also that while these translations both contain an error ("neemt . . . in de gaten"), this is not marked as an inconsistency, because it is shared between the translations.

(1) a. EN: The aunts criticise the {king, father}, and the man definitely observes the mayor.
b. NL: (. . .) en de man neemt zeker de burgemeester in de gaten.
c. NL: (. . .) en de man neemt de burgemeester zeker in de gaten.

Another commonly occurring case of rephrasing is one where the two translations include terms that are (nearly) **synonymous terms** in Dutch. Some examples are the translation of athlete ("sporter" vs "atleet"), wish ("wensen" vs "willen") and observe ("observeren" vs "waarnemen"). Some of them can appear in the same context but for others the two words would typically appear in different types of texts. For instance, the word "dokter" is used in more informal contexts than the word "arts" (both translations of "doctor"). Again, we could not identify an interpretable pattern for when the model emits one instead of the other – they were not understandably related to the modifications we made to the inputs.

F.2.2 Source ambiguities

An intriguing category that we had not anticipated were cases in which the source sentence contained ambiguities, such as **polysemous words** (e.g. "director" translated to "directeur", referring to the director of a company, and "regisseur", indicating the director of a movie). Other ambiguities encountered were **scope ambiguities**, that were particularly prominent for the systematicity test. In that test, we concatenate two sentences, and the ambiguity was often related to the verb in the first sentence – e.g. in Example (2):

(2) a. EN: The friend wishes that the {lawyers, directors} scream, and the victims (. . .)

While we intended this to be a conjunction of two independent sentences, there is also a reading where "wishes" takes scope over the entire second conjunct. In Dutch, those two cases are distinguishable because they trigger a different word order in the embedded clause (SOV), which is not grammatical for main clauses. Such scope changes often lead to very questionable interpretations of the English sentence, as is the case for Example (3):

(3) a. EN: The victims want that the {doctors, mayors} run, and the victims read an article about the case of a procedure which includes a repayment plan.
b. EN: The farmers think that the {butchers, mothers} laugh, and an error can only be seen

whenever we have a basic plan that is constantly compared to our real actions.

c. EN: The women wish that the {painters, victims} walk consciously, and every 2CV or Dyane can basically be used as a donor.

Interestingly, the models sometimes also changed the order in the relative clause when a scope change was not possible, for instance when the second conjunct was a question, or the verb in the first sentence did not allow to take scope over the second conjunct without the presence of the word "that". See Example (4). We underline the incorrect part of the translation, here and in erroneous examples that follow.

(4) a. EN: The victim observes the {leader, king}, and the fathers carefully avoid the president.
b. NL: Het slachtoffer observeert de leider en de vaders de president zorgvuldig vermijden.
c. NL: Het slachtoffer observeert de koning en de vaders vermijden voorzichtig de president.

These examples indicate that the interpretation of scope change might not be applicable here and that instead, the model is applying some heuristic where particular words trigger a relative clause order.

F.2.3 Target errors

In the category 'target errors', some of the errors can be easily traced to individual words, whereas others indicate overall misinterpretations of the input.

Single word errors Errors that consist of single words are caused by words that are either missing, wrongly translated or untranslated. Changes due to **missing words** can be very minor but nevertheless render one of the sentences ungrammatical (e.g. "De tante achter de truck bewonderde de directeur", correct, vs "De tante achter de truck bewonderde directeur", incorrect), or yield grammatical sentences that have a slightly different meaning (e.g. "de arts die yoghurt eet" vs "de arts die *de* yoghurt eet"). Missing words can also render translations both ungrammatical and semantically incorrect, which occurred mostly in case of missing nouns or verbs (e.g. "de bakker die ons herkent, merkt de koning op", correct, vs "de bakker die ons de koning herkent", incorrect).

We also encountered pairs where one translation contained **untranslated source words**. This happened with some of the words in our synthetic templates (e.g. "ooms"/"uncles", "butchers"/"slagers") but also with words from the natural sentences (e.g. "extrusion"/"extrusie", "soils"/"bodem"). These cases mark examples where local processing would have been helpful to the model: as evidenced by the alternative translation in the pair, the model does have access to the correct translation.

Thirdly, we observed cases of **mistranslated words**, where words unrelated to the change locus received a wrong translation in one of the two sentences but a correct one in the other, for example: "poets" being translated as "dichters" (correct) vs "de potten" (incorrect), "general" as "generaal" (correct) vs "wandeling" (incorrect), or "productform" as "productvorm" (correct) vs "productformulier" (incorrect).

Multi-word errors Other types of errors are less easily located to individual words but indicate an overall misinterpretation of the input, such as the **change in the tense** as displayed in Example (5), and the **change in agreement** displayed in Example (6). In these particular cases, the source of confusion is explainable: in the first case, the model is combining a present tense verb with a word-order that does not support that, even though such a word order does exist ("in het najaar van 2005 ... en komen er al snel een paar ..."). In the second case, "begrijpen" should agree with "schilder" but instead agrees with the word "doctors", much earlier in the sentence. In both of these cases, a more locally compositional approach to translating would have yielded correct translations.

(5) a. EN: (...) and in autumn 2005, five musicians join their forces and soon a couple of potential songs came into being in the rehearsal room.
b. NL: (...) in het najaar van 2005 voegen vijf muzikanten zich bij hun krachten en al snel kwamen er een paar potentiële nummers in de oefenruimte.
c. NL: (...) in het najaar van 2005 bundelen vijf muzikanten hun krachten en al snel komen er een paar potentiële nummers tot stand in de oefenruimte.

(6) a. EN: The doctors that laugh admire the {president, baker}, the painter that admires her understands the king.

b. NL: (...) de schilder die haar bewondert, begrijpen de koning.
c. NL: (...) de schilder die haar bewondert begrijpt de koning.

Finally, we would like to point out an error type that relates to the **semantic role assigned to agents**, and brings about a lot of other changes in the process. For instance, in Example (7), "the fathers" is removed from the main clause and moved into the relative clause, leaving the main clause without its direct object.

(7) a. EN: The group of painters behind the truck forgets the {president, friend} and an article about the previous EESC Opinion on alcohol related harm, which looked at f, is read by the fathers
b. NL: (...) en een artikel over het eerdere advies van het EESC over alcoholgerelateerde schade, die door de vaders wordt onderzocht, wordt gelezen.
c. NL: (...) en een artikel over het eerdere advies van het EESC over alcoholgerelateerde schade, die naar f uitkeek, wordt door de vaders gelezen.

F.2.4 Formatting

We marked inconsistencies as formatting changes if they were related to punctuation, capitalisation, hyphenation or differences in usage of spaces. In most cases, those cases were caused by commas: in one translation, a relative clause or two conjuncts were separated by a comma, whereas in the other translation the comma was left out. In the cases that were caused by spaces ("tumormassa" vs "tumor massa"), there is a slight difference in correctness: in Dutch, compound nouns are not separated by spaces. Given how minor these mistakes are, we did not mark them as errors. Example (6) above provides an example for inconsistent usage of commas. Formatting changes are far from the most frequent but they do become more prominent in models trained on larger training corpora.

F.2.5 Inconsistencies in synonym translations

The synonym errors are subdivided into cases where synonyms are simply translated differently (we observed this mostly for the models with larger training set sizes), cases where both translations were incorrect, cases in which only one translation is wrong, and cases in which one synonym was not translated but directly copied from the source. Sometimes, the changes were quite peculiar, to give some examples from our natural corpus:

(8) a. EN: The child admires the king that eats the {doughnut, donut}.
b. NL: Het kind bewondert de koning die de donut eet.
c. NL: Het kind bewondert de koning die de ezel eet.

(9) a. EN: - Yeah, a barbecue sauce {moustache, mustache} contest.
b. NL: - Ja, een barbecue [missing 'sauce'] met snor.
c. NL: - Ja, een barbeceu saus snor wedstrijd.

How often each of these errors occur depends on the synonym. Where some synonyms are more prone to being untranslated (like "ladybird" and "flautist"), some simply received many different correct translations (like "shopping trolley") yet others received errors very specific to the synonym (like "eggplant" being translated as "egg"+"plant", an interesting case because it reflects processing that is too local). It should be noted that for all synonyms – apart from the model with the small training dataset that cannot translate "flautist" and "ladybug" – we have observed correct translations, indicating that the models did in fact acquire their meaning.

Further, it should be noted that while our substitutivity experiment provides insight into how the model copes with individual synonyms, the majority of the inconsistencies observed were still common target errors, rephrasings, changes in formatting or the result of source-side ambiguities. It is vital here to stress that the types of rephrasings, however, did not appear related to the writing style of the sentence. For instance, considering that the synonym changes were related to British and American spelling, and occassionally changed the tone of the sentence (e.g. "aeroplane" could be considered more archaic compared to "airplane"), one could anticipate changes in word choice in Dutch reflecting this change of style. However, the inconsistencies were virtually indistinguishable from those annotated for systematicity.