Blackbird Language Matrices Tasks for Generalization

Paola Merlo, Chunyang Jiang, Giuseppe Samo, Vivi Nastase
University of Geneva
{Chunyang.Jiang, Paola.Merlo, Giuseppe.Samo}@unige.ch
vivi.a.nastase@gmail.com

Abstract

To develop a system with near-human language capabilities, we need to understand current systems' generalisation and compositional abilities. We approach this by generating compositional, structured data, inspired from visual intelligence tests, that depend on the problem-solvers being able to disentangle objects and their absolute and relative properties in a sequence of images. We design an analogous task and develop the corresponding datasets that capture specific linguistic phenomena and their properties. Solving each problem instance depends on detecting the relevant linguistic objects and generative rules of the problem. We propose two datasets modelling two linguistic phenomena – subject-verb agreement in French, and verb alternations in English. The datasets can be used to investigate how LLMs encode linguistic objects, such as phrases, their grammatical and semantic properties, such as number or semantic role, and how such information is combined to correctly solve each problem. Specifically generated error types help investigate the behaviour of the system, which important information it is able to detect, and which structures mislead it.

1 Motivation

The current reported success of large language models (LLMs) is based on computationally expensive algorithms and large amounts of data that are available for only a few, non-representative languages. Such data may also contain biases and imbalances, and its sheer size prevents curation. To be able to build robust models that can learn better from manageable sized data, we need to understand the current systems' generalisation and compositional abilities.

We argue that a system with high language competence and performance, that is able to learn from small amounts of data, and is cross-linguistically valid, should capture the three fundamental properties of human language: (i) human language is described by several abstract levels of representations (e.g. morphological, phonological, syntactic, semantic), mapped onto each other by complex many-to-many rules; (ii) it is compositional; (iii) it is structured.

For GenBench, we propose several datasets under the same umbrella, as they have the same format, but encode different linguistic phenomena, each in a different language – subject verb agreement in French, verb alternations in English. They can be used separately, or in combination, to explore the properties and the generalisation abilities of a LLM in various ways.

- Test whether sentence representations encode the targeted linguistic information.
- Test generalisation when data has different levels of lexical variation.
- Providing probes into how sentence representations encode the targeted information – by studying different minimal architectures that aim to find patterns in pretrained sentence representations.[1]
- Providing cross-linguistic and multi-task probes for detecting how sentence representations encode different kinds of targeted linguistic information. The fact that our datasets have the same structure allows for a variety of experimental set-ups to probe how sentence embeddings encode different linguistic phenomena across different languages.

Additional datasets are in development, thus expanding the scope of the exploration. With respect to the workshop aims, our motivations are as follows.

[1] By *minimal* we mean the least complex architectures that could be used to discover patterns in the input data.

Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pages 163–172
December 6, 2023 ©2023 Association for Computational Linguistics

Cognitive : explore how specific linguistic information is encoded in pretrained sentence representation, and determine whether, or to what degree, we can identify symbolic structures and compositional elements within these representations.

Intrinsic : explore whether pretrained LLMs have learned language representations whose properties can be mapped onto those proposed in linguistics, and whether the tasks we propose are solved through identifiable rules.

While the proposed datasets are presented as a diagnostic tool – to detect patterns that encode linguistic rules and phenomena – we envision that they could also be used to bring such patterns to the fore, through fine-tuning pretrained sentence embeddings, thus pushing continuous distributed representations towards more symbolic, and interpretable, ones.

2 Blackbird Language Matrices

The design of our datasets were inspired by Raven Progressive Matrices (RPMs) (Raven, 1938), an example of which is presented in Figure 1.

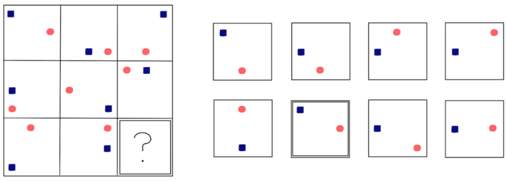


Figure 1: An example Raven's progressive matrix (best seen in colour). The matrix is constructed according to two rules: (i) the red dot moves one place clockwise when traversing the matrix left to right; (ii) the blue square moves one place anticlockwise when traversing the matrix top to bottom. The task consists in finding the tile in the answer set that correctly completes the sequence, indicated with a double border.

RPMs are used in visual IQ tests, as they rely on problem-solvers identifying elements and their attributes such as position, shape, colour and size, and their absolute and relative properties (for instance, how their positions change relative to each other throughout the matrix of images). Analogously, in language, elements correspond to phrase types, attributes correspond to grammatical gender or number, or specific semantic properties, and their connective properties are the relative positions within a syntactic structure or the mapping across levels of representations.

2.1 The Blackbird Language Matrices (BLM) task

Merlo et al. (2022); Merlo (2023) describe the Blackbird Language Matrices (BLM) task. A targeted linguistic phenomenon is presented in the form of a set of sentences that have both syntagmatic and paradigmatic relations. This way, like in the RPM visual version, they give rise to a matrix structure. The language matrices manipulate phrases, dependencies in the syntactic tree, and lexical, grammatical and semantic attributes between connected elements of a sentence and across sentences.

A BLM task comprises a context and an answer set: the context C is a sequence of sentences that share the targeted grammatical phenomenon, but differ in other relevant aspects. BLMs are multiple-choice problems, and each context is paired with a set of candidate answers W. The incorrect answers are built by corrupting the generating rules of the context sequence. This contrastive set up enables targeted error analyses and provides information on structures learned and the type of mistakes a system is prone to. More formally, a BLM task, problem, and matrix can be defined as follows.

BLM TASK: Find $(w_c \in W)$ given C,

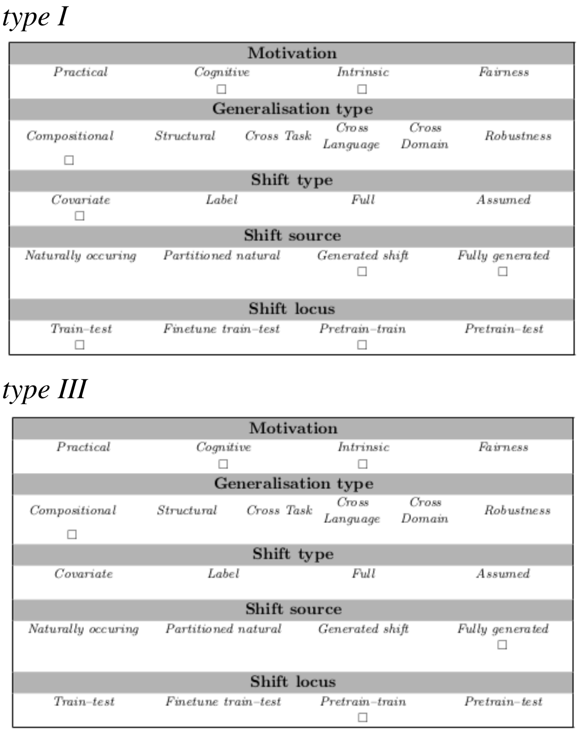
given a 4-tuple (LP, C, W, w_c), where LP is the definition of the linguistic grammatical phenomenon, C is the corresponding context matrices, W is the answer set, and w_c is the selected item of W that is correct.

BLM PROBLEM: A *BLM problem* is a tuple(LP, C, W, Aug). It is an instance of a BLM task, where Aug is the augmentation method for the matrices.

BLM MATRIX: A *BLM matrix* is a tuple (S, R, T) s.t. S is the shape of the matrix, R are the relational operators that connect the items of the matrix, T is the set of items of the matrix.

2.2 The BLM Datasets

We propose two datasets encoding two different linguistic phenomena, in different languages: subject-verb agreement in French, and verb alternations in English. We submit to the GenBench task two variations for each dataset: one where each problem in the training data consists of a sequence of sentences with minimal lexical variation (*type I*), and one where the lexical variation is maximal (*type III*). Figure 2 shows the evaluation cards for the two types of datasets (with training with minimal lexical variation, and the training and test data sampled from the same population of automatically generated instances). Table 1 shows the dataset statistics.



type I

Motivation					
Practical	Cognitive □	Intrinsic □	Fairness		
Generalisation type					
Compositional □	Structural	Cross Task	Cross Language	Cross Domain	Robustness
Shift type					
Covariate □	Label	Full	Assumed		
Shift source					
Naturally occuring	Partitioned natural	Generated shift □	Fully generated □		
Shift locus					
Train–test □	Finetune train–test	Pretrain–train □	Pretrain–test		

type III

Motivation					
Practical	Cognitive □	Intrinsic □	Fairness		
Generalisation type					
Compositional □	Structural	Cross Task	Cross Language	Cross Domain	Robustness
Shift type					
Covariate	Label	Full	Assumed		
Shift source					
Naturally occuring	Partitioned natural	Generated shift	Fully generated □		
Shift locus					
Train–test	Finetune train–test	Pretrain–train □	Pretrain–test		

Figure 2: Evaluation cards *type I* (top) and *type III* (bottom).

The shift and generalisation types are as follows.

Shift source : *fully generated* – BLM-AgrF has been automatically generated starting from manually selected seeds and provided templates. + *generated shift*: *type I* variation contains training data sampled from a different distribution than the test data.

Shift type : *covariate*: for *type I* there is a covariate shift between training and testing input data.

Shift locus : *pretrained-trained* – the datasets are designed to make use, as input, of the representations produced by pretrained LLMs, and use them in a novel task. + *train-test* – for the *type I* variations.

Generalisation We aim for *compositional* generalisation, by proposing a dataset that can be used to probe whether different linguistic objects, their properties, and the rules through which they combine are identifiable in pretrained sentence representations.

2.2.1 BLM-AgrF: Subject-verb agreement (in French)

Subject-verb agreement is often used to test the syntactic abilities of deep neural networks (Linzen et al., 2016; Gulordava et al., 2018; Goldberg, 2019; Linzen and Baroni, 2021). While theoretically simple, it can have several complicating factors, such as intervening elements between nouns and the verb, which can interfere with the proper matching of the agreement features.

	Context			
1	Le vase	avec la fleur		est cassé.
2	Les vases	avec la fleur		sont cassés.
3	Le vase	avec les fleurs		est cassé.
4	Les vases	avec les fleurs		sont cassés.
5	Le vase	avec la fleur	du jardin	est cassé.
6	Les vases	avec la fleur	du jardin	sont cassés.
7	Le vase	avec les fleurs	du jardin	est cassé.
8	???			

	Answer set	
1	Le vase avec la fleur et le jardin est cassé.	coord
2	**Les vases avec les fleurs du jardin sont cassés.**	correct
3	Le vase avec la fleur est cassé.	WNA
4	Le vase avec la fleur du jardin sont cassés.	AE
5	Les vases avec les fleurs du jardin sont cassés.	WN1
6	Les vases avec les fleurs des jardins sont cassés.	WN2

Figure 3: BLM instances for verb-subject agreement, with two attractors (*fleur* (flower), *jardin* (garden)), with candidate answer set. WNA=wrong number of attractors, AE=agreement error, WN1=wrong nr. for 1^{st} attractor noun (N1), WN2=wrong nr. for 2^{nd} attractor noun (N2)

In BLM-AgrF (An et al., 2023),[2] a BLM problem for subject-verb agreement consists of a context set of seven sentences that share the subject-verb agreement phenomenon, but differ in other aspects – e.g. number of intervening noun phrases between the subject and the verb, called attractors because they can interfere with the agreement, different grammatical numbers for these attractors, and different clause structures. Each context is paired with a set of candidate answers. The answer sets contain minimally contrastive examples built by corrupting some of the generating rules. This helps investigate the kind of information and structure learned, by error analysis. An example is given in Figure 3.

[2] The names of the datasets are composed of a descriptor of the grammatical phenomenon (usually three letters) and the initial of the language (Agr = Agreement; F = French).

Example of context	
1	The girl sprayed the wall with paint.
2	Paint was sprayed by the girl
3	Paint was sprayed onto the wall by the girl
4	Paint was sprayed onto the wall
5	The wall was sprayed by the girl
6	The wall was sprayed with the paint by the girl
7	The wall was sprayed with paint
8	???

Example of answers	
The girl sprayed paint onto the wall	Correct
The girl was sprayed paint onto the wall	AgentAct
The girl sprayed paint the wall	Alt1
The girl sprayed with paint onto the wall	Alt2
The girl sprayed paint for the room	NoEmb
The girl sprayed paint under the wall	LexPrep
Paint sprayed the girl onto the wall	SSM
The wall sprayed the girl with paint	SSM
Paint sprayed the wall with the girl	AASSM

Figure 4: Verb alternations (ALT-ATL): a minimally lexicalised data instance. The labels indicate which (sub)rules are corrupted to create the error. See text for explanation.

Example of context	
1	The girl sprayed paint onto the wall.
2	Paint was sprayed by the girl
3	Paint was sprayed onto the wall by the girl
4	Paint was sprayed onto the wall
5	The wall was sprayed by the girl
6	The wall was sprayed with the paint by the girl
7	The wall was sprayed with paint
8	???

Example of answers	
The girl sprayed the wall with paint	Correct
The girl was sprayed the wall with paint	AgentAct
The girl sprayed the wall the paint	Alt1
The girl sprayed onto the wall with paint	Alt2
The girl sprayed the wall of the room	NoEmb
The girl sprayed the wall under the paint	LexPrep
The wall sprayed the girl with the paint	SSM
Paint sprayed the girl onto the wall	SSM
The wall sprayed the paint with the girl	AASSM

Figure 5: Verb alternation (ATL-ALT), a minimally lexicalised data instance. The labels indicate which (sub)rules are corrupted to create the error. See text for explanation.

2.2.2 BLM-s/lE: verb alternations (in English)

The study of the argument-structure properties of verbs and semantic role assignments is also a testbed for the core syntactic and semantic abilities of neural networks (Kann et al., 2019; Yi et al., 2022). Specifically, Yi et al. (2022) demonstrates that transformers can encode information on the two alternants of the well-studied *spray-load* alternation (Levin, 1993).

The BLM dataset for investigating the encoding of alternation properties is BLM-s/lE (Samo et al., 2023).[3] A naturally occurring example for each verb was extracted from the Spike Amazon subcorpus (Shlain et al., 2020), adopted as seeds for data-augmentation with a fill-mask task. Details are given in (Samo et al., 2023). A BLM s/lE matrix consists of a context set comprising one alternant (e.g. *The girl sprayed the wall with paint*) of the *spray-load* alternation and other sentences that provide the syntactic properties of the arguments of the alternation (e.g. passivization strategies). The target sentence is the other alternant (in our case, *The girl sprayed paint onto the wall*) to be chosen from an answer set of superficially minimally, but, syntactically and semantically, deeply different candidates. An example matrix is shown in Figure 5. We created two templates, one for each of the two alternates. One group has the alternant Agent-Locative-Theme (hencefort ALT, e.g. *The girl sprayed the wall with paint)* in the context and the correct answer is the alternant whose configuration is Agent-Theme-Locative (henceforth ATL, e.g. *The girl sprayed paint onto the wall*). ALT-ATL data is the data produced from the matrix in Figure 5.[4]

The answer set is contrastive – see caption of Figure 5. The answer labelled as AgentAct minimally deviates from the correct answer, since the verb is inflected in a passive mood; in Alt errors, the verb of the alternate is followed by two NPs and one PPs; in NoEmb errors, the PP is syntactically embedded in the NP; LexPrep errors involve a preposition which does not grammatically belong to the alternation. Finally, violations of the syntax-semantic mapping (SSM1 and SSM2) and simultaneous violations of AgentAct and SSM (AASSM) involve reorderings of the lexical constituents and functional elements (e.g. prepositions).

3 Benchmarking

We used two baselines to benchmark the proposed datasets. They are designed to test whether we can access the relevant information for the targeted phenomena in a given BLM task, in transformer-based sentence representations. Figure 6 shows

[3] The name follows our convention: s/l = spray/load; E = English. The dataset is created on the basis of a class of 30 verbs belonging to the same class of *spray* and *load* in VerbNet (Schuler 2005).

[4] The name of the data subset, ALT-ATL, is transparent towards this logic: ALT is given in the context set, ATL is the correct answer. The second group's template (ATL-ALT) is the converse.

the general process flow. The two baselines are a feed-forward neural network (FFNN) and a convolutional neural network (CNN).

The FFNN baseline is a three-layer feed-forward neural network. It transforms the context C of a BLM instance into a 1D-tensor which is a concatenation of the representation of each sentence. This is passed through three fully-connected layers. The output is a vector that we take to represent the embedding of the answer sentence. This architecture allows the system to find patterns within and across sentences through the nodes in the successive layers.

The CNN baseline consists of three convolutional steps, followed by a linear layer to compress the output to the desired dimensions. The input consists of a stack of context sentence representations. This setup finds localized patterns within sentence representation and across the sequence of sentences.

The output of the two networks is the same – a vector representing the sentence embedding of the correct answer. The learning objective is to maximize the probability of the correct answer from the candidate answer set. Because the incorrect answers in the answer set are specifically designed to be minimally different from the correct answer, we implement the objective through the max-margin loss function. This function combines the distances between the predicted answer and the correct and erroneous ones. We first compute a score for the embedding e_j of each candidate answer a_j in the answer set $\mathcal{A}$ with respect to the predicted sentence embedding e_{pred} as the cosine of the angle between the respective vectors:

$$score(e_j, e_{pred}) = cos(e_j, e_{pred})$$

The loss uses the max-margin between the score for the correct answer e_c and for each of the incorrect answers e_i:

$$loss = \sum_{e_i} [1 - score(e_c, e_{pred}) + score(e_i, e_{pred})]^+$$

At prediction time, we take the answer with the highest $score$ value from a candidate set as the correct answer.

4 Results and Error Analysis

The train/test data splits are presented in Table 1. As the task is set-up as multiple choice, we measure the results in terms of F1 scores for identifying the correct answer. The results below also show the performance on the test set for varying amounts of training data to show their impact, and compares two types of pretrained sentence embeddings – RoBERTa (Liu et al., 2019) and Electra (Clark et al., 2020).

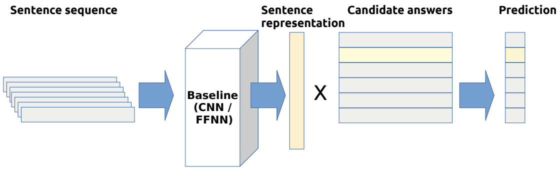


Figure 6: Illustration of the baseline setup experiments.

Datasets		*type I*	*type III*
BLM-AgrF		2073/3840	34650/3840
BLM-s/lE	ALT-ATL	3375/1500	13500/1500
	ATL-ALT	3375/1500	13500/1500

Table 1: Datasets statistics in terms of train/test counts.

4.1 Varying the training data

The results below show the performance on the test set in terms of F1 averages over five runs for each of the two datasets, for RoBERTa and Electra sentence embeddings.[5] The plots in Figure 7 show the results for the *type III* dataset variations (with train and test data sampled from the same population with maximal lexical variation), and the results for the *type I* dataset variations (with training data with minimal lexical variation within an instance). The results obtained with the overall baseline system (FFNN with Electra sentence embeddings) are shown in the tables in the left column.

The results shown in Figure 7 reveal interesting distinguishing properties of the two tasks. For the subject-verb agreement, which is a syntactic task, both types of sentence embeddings lead to similar results when using the FFNN system. Instead, a difference arises across architectures. The fact that the CNN leads to lower performance indicates that it finds more localised patterns and it also indicates that patterns capturing subject-verb agreement are more spread throughout the sentence embeddings. For the verb alternation task, which has a strong semantic component, the embedding type makes more of a difference than the system used to detect patterns. Electra seems to encode verb semantics better for this task, as Yi et al. (2022) also note. Because both the FFNN and the CNN detect successfully these patterns, this indicates that patterns

[5] For all sets of five runs the standard deviation was less than $1e - 10$, so it is not included in the tables.

Results for best baseline (Electra + FFNN)

tr+dev (80:20)	*type I*	*type III*
50	0.293	0.332
100	0.345	0.374
200	0.451	0.402
500	0.607	0.468
1000	0.661	0.622
1500	0.676	0.711
2000	0.702	0.741

tr+dev (80:20)	*type I*	*type III*
50	0.601	0.795
100	0.676	0.801
200	0.732	0.814
500	0.729	0.882
1000	0.795	0.896
1500	0.795	0.943
2000	0.798	0.938
3000	0.803	0.929

tr+dev (80:20)	*type I*	*type III*
50	0.600	0.731
100	0.665	0.806
200	0.726	0.848
500	0.809	0.876
1000	0.793	0.887
1500	0.800	0.915
2000	0.809	0.903
3000	0.794	0.914

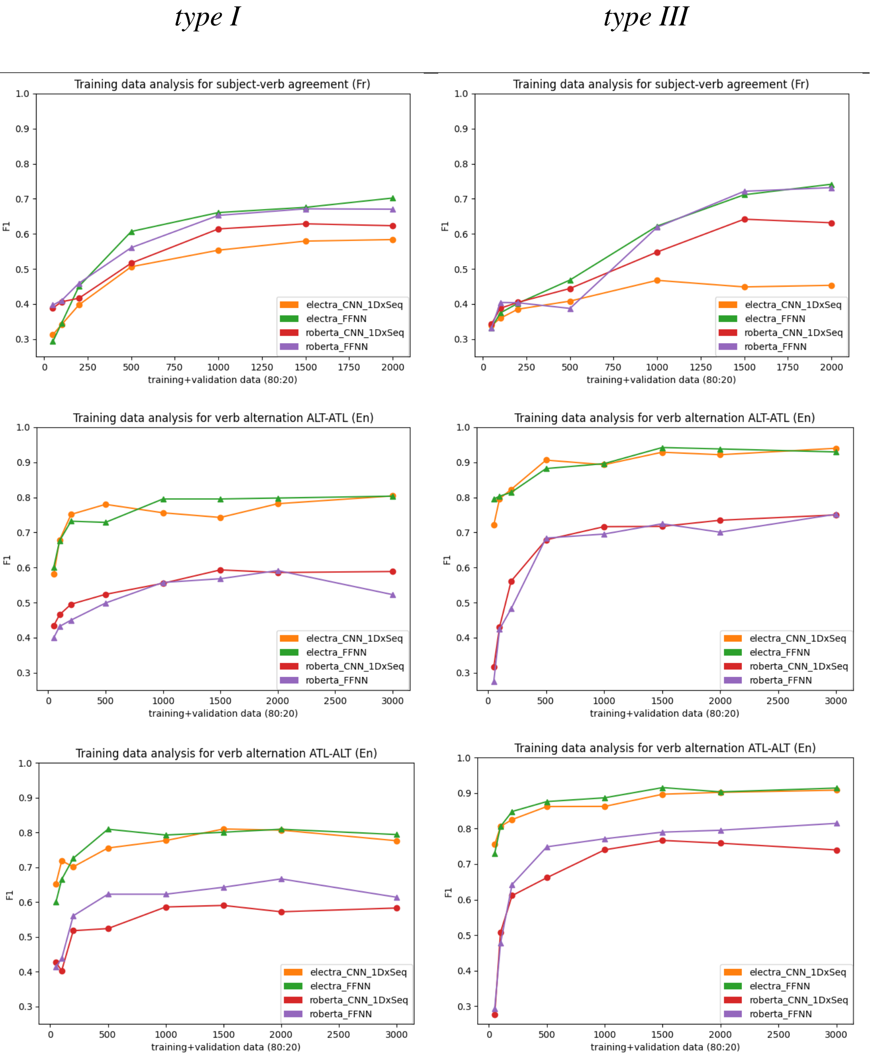


Figure 7: Result plots in terms of F1 averages over five runs, when using the two baselines and RoBERTa and Electra sentence embeddings, and numeric results in the tables for the best combination: Electra with FFNN baseline system

encoding verb alternations are more localised.

Having training data with minimal lexical variation makes the targeted pattern more obvious. On the other hand, it may provide shallow indicators that can confound the system. Comparing the results on *type I* and *type III*, we note that there is a drop of about 0.1 in the F-score for all settings, although for the subject-verb agreement this is lower (0.04). This is probably not surprising and underlines the more structural nature of the subject-verb agreement problem, where lexical variation does not detract from the number agreement pattern. For the verb alternation the drop is higher. This may suggest that since the task is more semantic in nature, variation in the lexical material of the sentence shifts the underlying patterns. Some transformation of the sentence representations may make such patterns more obvious, and separate them from the lexical signal. However, the performance is still high, even with a smaller amount of training data, indicating that the signal that encodes the *spray-load* alternation is strong in the sentence embeddings.

4.2 Error Analysis

Error analysis on the best baseline – RoBERTa sentence embeddings with the CNN system – is given in Figure 8. The upper panel refers to BLM-AgrF, the lower panel provides information about the errors within the ALT-ATL dataset BLM-s/lE.

In both datasets, we observe clear trends. First, more data reduces errors roughly uniformly. Minimal variation is observed in BLM-AgrF, where

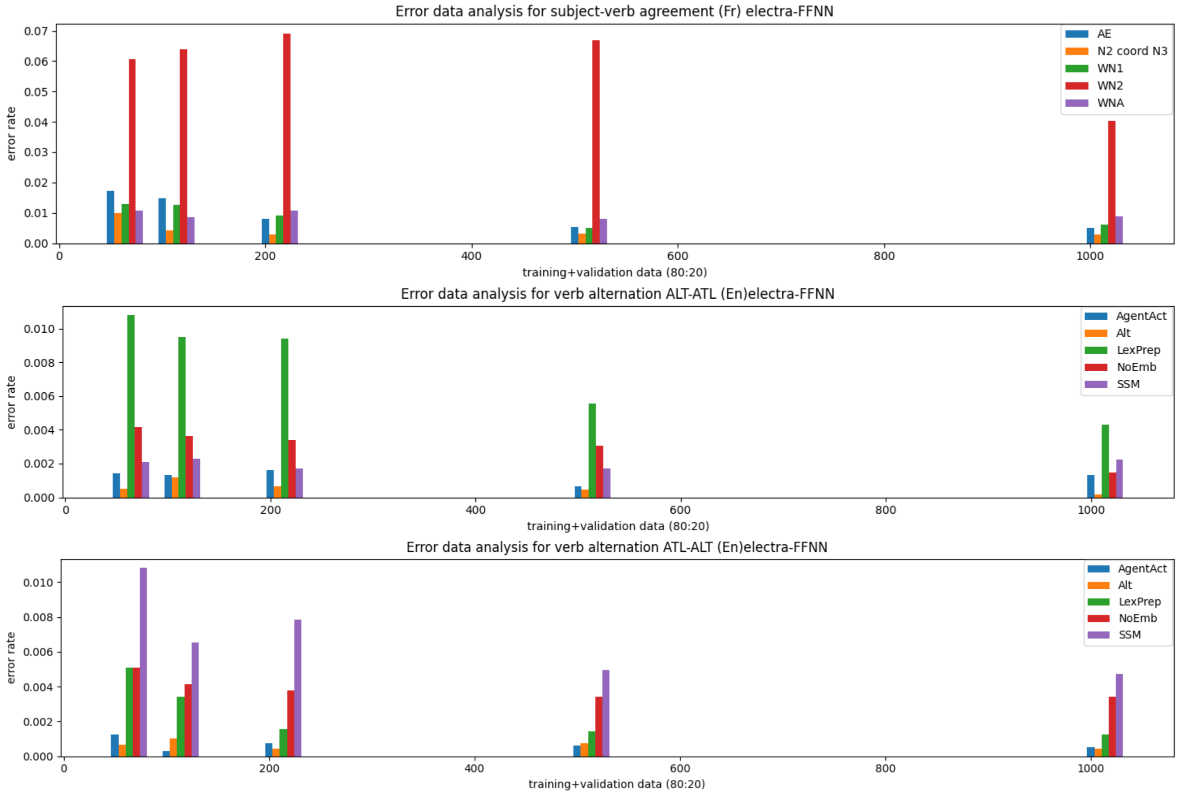


Figure 8: Error analysis (averages over 5 runs) for varying amounts of training data, for the best performing baseline: RoBERTa sentence embeddings with the CNN system.

WN2 (wrong number for 2^{nd} attractor noun) remains the most frequent error across size of the training data. This error has two interesting characteristics: it is the most frequent error also for human speakers, and also it is exhibited by sentences that are grammatical, but do not respect the global pattern of the matrix.

Conversely, a conspicuous trend is distinctly observable within the BLM-s/lE dataset (ALT-ATL). The distribution of mistake types concentrates on lexical mistakes of functional elements (prepositions) with small training sets. However, as the dataset size increases, the SSM error increases proportionally. This error is associated with the semantic properties of the alternation, specifically the semantic roles of the arguments of the verb.

5 Related work

The GenBench taxonomy differs sometimes from the meaning of some existing terms elsewhere referring to generalisation. In situating our work in comparison to other related work, we reason based on the actual nature of the generalisation being sought rather than the terminology.

Our closest related dataset in spirit, in terms of motivation and goals, is the COGS dataset (Kim and Linzen, 2020). It is also different from our dataset in implementation. The COGS dataset aims at providing out-of-distribution test cases to test compositionality of structure and meaning. To this goal, a training set is generated with a CFG and parallel lambda-expressions and a test set with a different CFG, specifically designed to exhibit testing constructions that are previously unseen as such and whose solution requires compositional generalisation of components seen at training.

These unseen constructions comprise both structural and lexical generalisation: the former aiming to test the ability to create new structures from existing parts, the latter to test ability to adapt existing structures to novel content.

While our dataset strives for similar goals, the way to go about it is different in one relevant respect. The COGS dataset determines by design the combinatorics that the network needs to find, imposing therefore preexisting hard independence assumptions generated by a CFG in the test set. These pre-existing discrete rules of combination must be discovered to find a correct parsing solution.

Our approach is more in the spirit of hidden representation learning. We do not require that the network has explicitly learnt new generative ways of combining elements. But we encourage representations that learn soft constraints, in the form of disentangled representations that correspond to the generative underlying factors. Beside testing soft constraints on structural generalisation, we also provide tests of lexical generalisation (through the different types of lexical variability in the matrix).

In this respect, the datasets provided here are related to those used in the literature on disentanglement in computer vision. For example, van Steenkiste et al. (2020) developed a dataset for computer vision similar to RPMs. They evaluate the usefulness of the representations learned for abstract reasoning. They note that learning disentangled representations leads to faster few-shot learning. Also, recently Zheng and Lapata (2022) propose a different method for disentangling relations expressed in a sentence which may share arguments. This is implemented as an extension to sequence-to-sequence (seq2seq) models, where at each decoding step the source input is re-encoded by conditioning the source representations on the newly decoded target context. These specialized representations make it easier for the encoder to exploit relevant-only information for each prediction.

With the appropriate dataset, such approaches can be used to probe the abilities of pretrained LLMs. The datasets we propose in this paper have the necessary properties: they focus on specific linguistic phenomena, they display lexical and structural variation, and include known confounding factors for the targered phenomena. They are, then, close to the work that investigates network representations. For example, Lasri et al. (2022) focus on how BERT encodes grammatical number in English and how this information is used for performing number agreement. The focus is on word embeddings and quantifying how much number information they encode at various layers of the BERT architecture. Using a combination of probing approaches, they discover that subjects and predicates embeddings do encode number information, but at different layers. Further investigations into where and how the number information is shared reveals that number information is not directly shared, but rather passed through intermediate tokens. The study of the argument-structure properties of verbs and semantic role assignments is also a test-bed for the core syntactic and semantic abilities of neural networks (Kann et al., 2019; Yi et al., 2022). In particular, Yi et al. (2022) demonstrates that transformers can encode information on the two alternants of the *spray-load* alternation.

6 Conclusions

In this paper, we describe an approach to generate compositional, structured data, inspired from visual intelligence tests. We presented two datasets, each focused on a different linguistic phenomenon, and in a different language. Solving each problem instance depends on the system detecting the relevant linguistic objects, and their absolute and relative properties. These datasets can be used to investigate whether this type of information can be detected in, and whether it is used by, pretrained LLMs. Because the datasets are formatted in the same way, they can be used separately, or in various combinations, to test cross-task and cross-language model properties. Additional such datasets are under development, thus potentially expanding the scope of the exploration.

Further experiments are also ongoing. One of our goals is to understand how information is encoded in pretrained transformer-based sentence embeddings. We investigate whether there are patterns within sentence representations that reveal specific linguistic phenomena. Towards this end, we have developed architectures designed to discover such patterns that can be applied successfully, without adaptation (in terms of architecture or hyperparameters) to different problems in different languages. This provides insight into how transformers encode sentence-level information.

Limitations

The approach is evaluated on a limited range of syntactic phenomena and models. Expanding the scope could better demonstrate the general utility. In particular, we would like to expand in many directions: (i) the structures that are tried in the different test sets; (ii) the different phenomena under study; (iii) the complexity of the matrices, which can be made progressively harder by combining linguistic phenomena in a single matrix. Finally, we need to tackle the complex problem of how to generate more naturally structured data, while retaining the controllable nature of synthetic, experimental data.

Ethics Statement

To the best of our knowledge, this paper raises no ethics concerns.

Acknowledgements

We gratefully acknowledge the partial support of this work by the Swiss National Science Foundation, through grants #51NF40_180888 (NCCR Evolving Language) and SNF Advanced grant TMAG-1_209426 to PM.

References

Aixiu An, Chunyang Jiang, Maria A. Rodriguez, Vivi Nastase, and Paola Merlo. 2023. BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1363–1374, Dubrovnik, Croatia. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 287–297.

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, Online. Association for Computational Linguistics.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paola Merlo. 2023. Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Motivations and formal specifications. *ArXiv*, cs.CL 2306.11444.

Paola Merlo, Aixiu An, and Maria A. Rodriguez. 2022. Blackbird's language matrices (BLMs): a new benchmark to investigate disentangled generalisation in neural networks. *ArXiv*, cs.CL 2205.10866.

John C. Raven. 1938. Standardization of progressive matrices. *British Journal of Medical Psychology*, 19:137–150.

Giuseppe Samo, Vivi Nastase, Chunyang Jiang, and Paola Merlo. 2023. BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. 2020. Syntactic search by example. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 17–23, Online. Association for Computational Linguistics.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. 2020. Are disentangled representations helpful for abstract visual reasoning? In *NeurIPS 2019*.

David Yi, James Bruno, Jiayu Han, Peter Zukerman, and Shane Steinert-Threlkeld. 2022. Probing for understanding of English verb classes and alternations in large pre-trained language models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 142–152, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hao Zheng and Mirella Lapata. 2022. Disentangled sequence to sequence learning for compositional generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4256–4268, Dublin, Ireland. Association for Computational Linguistics.