Building structured synthetic datasets: The case of Blackbird Language Matrices (BLMs)

Paola Merlo, Giuseppe Samo, Vivi Nastase* and Chunyang Jiang

Department of Linguistics, University of Geneva, Switzerland

Abstract

Our goal is to investigate, ultimately to enhance, to what degree existing LLM learn disentangled rule-based, compositional linguistic representations. We take the approach of developing curated synthetic data on a large scale, with specific properties, and using them to study sentence representations built using pretrained language models. Inspired by IQ tests, we develop a new multiple-choice task. Finding a solution to this task requires a system detecting complex linguistic patterns and paradigms in text representations. We present formal specifications of this task, illustrate it with two problems and present their benchmarking results.

Il nostro obiettivo è indagare, allo scopo di migliorare, quanto gli LLM esistenti apprendano rappresentazioni linguistiche composte, basate su regole districate. Il nostro approccio consiste nello sviluppare dati sintetici curati su larga scala, con proprietà specifiche, e nell'utilizzarli per studiare le rappresentazioni di frasi costruite con modelli linguistici pre-addestrati. Ispirandoci ai test del QI, abbiamo sviluppato un nuovo task a scelta multipla. Trovare la soluzione di questo task richiede che il sistema individui schemi e paradigmi linguistici complessi nelle rappresentazioni testuali. Presentiamo le specifiche formali di questo task, lo illustriamo con due problemi e presentiamo i risultati del benchmarking.

Keywords

synthetic structured data, formal definitions of grammatical phenomena, diagnostic studies of deep learning models

1. Introduction

Current consensus about LLM, and NNs in general, is that to reach better, possibly human-like, abilities, we need to develop tasks and data that help us understand their current generalisation abilities and help us train or tune them towards more complex and compositional skills.

Humans are good generalizers. A large body of literature has demonstrated that the human mind is predisposed to generate rules from data and combine these rules, in ways that have been argued to be distinct from the patterns of activation of neural networks [1, 2, 3]. One possible approach to develop more robust methods, then, is to drive the network to learn disentangled decompositions of complex observations and learn underlying regularities [4].

Let's look at an illustrative example of what complex decomposition of covert rules would be necessary. Consider complex argument structure relations in the lexicon: for example, the *Spray/load alternation* in English, shown in (1).

This alternation applies to verbs such as *spray, paint, spread, fill, stuff* and *load*, that describe covering surfaces or filling volumes [5, 6]. They occur in two subcategorisation frames, related to each other in a regular way: the object of the preposition *with* is the subject of the *onto* frame, while the object of the *onto* prepositional phrase is the subject of the *with* frame.

(1)	John	loaded	the truck	with hay.
	AGENT		LOCATIVE	THEME
	John	loaded	hay	onto the truck.
	AGENT		THEME	LOCATIVE

To learn the structure of such a complex alternation automatically, a neural network must be able to identify the elements manipulated by the alternation, and their relevant attributes, and recognize the operations that manipulate these objects, across more than one sentence.

To study what factors lead to learning more disentangled linguistic representations —representations that reflect the underlying linguistic rules of grammar— we take the approach of developing curated synthetic data on a large scale, building diagnostic models from pretrained representations of these data and investigating the models' behaviour. To this end, we develop a new linguistic task, inspired by the IQ test RPM (Raven 1938), which we call Blackbird Language Matrices (BLMs). BLMs define a prediction task to learn complex linguistic patterns and paradigms [7, 8].

In this paper, we present precise formal specifications of the BLM task, illustrate it with the instantiations of two BLM problems and their benchmarking results. This

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

*Corresponding author.

Paola.Merlo@unige.ch (P. Merlo); Giuseppe.Samo@unige.ch (G. Samo); vivi.a.nastase@gmail.com (V. Nastase); chunyang.jiang42@gmail.com (C. Jiang)

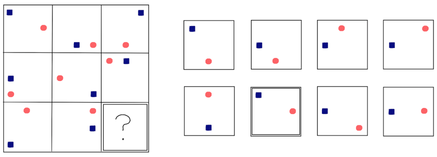

Figure 1: Example of progressive matrice in the visual world. The multiple-choice task is to determine the missing element in a visual pattern. The matrix is constructed according to two rules (see text for explanation). Identifying these rules leads to the correct answer (marked by double edges).

shows that the general formalism can be used to generate datasets with the same format, and similar specification. Expanding the covered phenomena can thus be done systematically, allowing for studies that combine or work across multiple phenomena and languages [9]. We believe this task takes us closer to investigations of human linguistic intelligence.

2. RPMs and BLMs

Raven's progressive matrices are IQ tests consisting of a sequence of images, called the *context*, connected in a logical sequence by underlying generative rules [10]. The task is to determine the missing element in this visual sequence, the *answer*. An instance is shown in Figure 1: given a matrix (left), choose the last element of the matrix from given options. The matrices are built according to generative rules that span the whole sequence of stimuli and the answers are constructed to be similar enough that the solution can be found only if the rules are identified correctly. For example in Figure 1, the matrix is constructed according to two rules: Rule 1: row-wise, from left to right, the red dot moves one place clockwise each time. Rule 2: column-wise, from top to bottom, the blue square moves one place anticlockwise each time. Identifying these rules leads to the correct answer, the only cell that continues the generative rules correctly.

A similar task has been developed called Blackbird Language Matrices (BLMs) [7, 8, 11] for linguistic problems, as given in Figure 2, which illustrates the template of a BLM agreement matrix. As can be seen, the agreement rules are implicitly expressed by patterns in the sentence, with or without intervening attractor elements, and alternate in a combinatorial pattern across the sentences (shown in colour), so that only one answer concludes the sequence.

3. Formal Specifications of BLMs

We define here the new Blackbird's Language Matrices (BLMs) task and data format.

Context				
1	NP-sing	PP1-sing		VP-sing
2	NP-plur	PP1-sing		VP-plur
3	NP-sing	PP1-plur		VP-sing
4	NP-plur	PP1-plur		VP-plur
5	NP-sing	PP1-sing	PP2-sing	VP-sing
6	NP-plur	PP1-sing	PP2-sing	VP-plur
7	NP-sing	PP1-plur	PP2-sing	VP-sing
8	???			

Answer		
1	NP-sing PP1-sing et NP2 VP-sing	Coord
2	**NP-plur PP1-plur PP2-sing VP-plur**	correct
3	NP-sing PP-sing VP-sing	WNA
4	NP-sing PP1-sing PP2-sing VP-plur	AE
5	NP-plur PP1-sing PP1-sing VP-plur	WN1
6	NP-plur PP1-plur PP2-plur VP-plur	WN2

Figure 2: BLM instances for verb-subject agreement, with two attractors. 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).

Subject-verb agreement
E: the subject and the verb match
in agreement features.
I: occurs independently of distance
between subject and verb.

Spray/Load alternation:
E: the object of alternant 1 becomes PP(P) in alt. 2.
the object of alternant 2 becomes PP(P) in alt. 1.
the preposition with of alternant 1 becomes
the preposition onto in alternant 2.
I: Expression of thematic roles and argument structure:
Object of Alt1 is Locative, PP of Alt1 is Instrumental
Object of Alt2 is Theme, PP of Alt2 is Locative

Figure 3: Examples of formal definitions of E and I for two LPs. The lexical expression of preposition P is considered an attribute.

Definition Let a 4-tuple (LP, C, W, w_c) be given, 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 correct item of W.

The BLM *task* can be defined as the instruction:

find $(w_c \in W)$ given C.

A BLM *problem* (LP, C, W, Aug) is an instance of a BLM task, where Aug is the augmentation method for the matrices. We describe all components in the next sections.

3.1. Defining the linguistic phenomenon

The first step in the definition of the problem consists in formally defining the linguistic grammatical phenomenon as a paradigm.

Definition Let a linguistic phenomenon LP be given. LP is exhaustively defined by a grammar $G_{LP} = (O, A, E, I, L)$ s.t.

O is the set of objects

A is the set of attributes of the objects in O

E is the set of external observed rules

I is the set of unobserved internal rules

L is the lexicon of objects in O, attributes in A, and operators in $E \cup I$.

For example, as shown in the example Figure 3, in the subject-verb agreement phenomenon, the agreement rule is the primary production in E, while the fact that agreement can occur independently of the distance of the elements expresses the fact that agreement applies to structural representations, a rule in I. Sometimes, but not always, I acts as a confusing factor.

Rules are triples of objects (shown in red), attributes (in green) and operations (in blue). Objects are usually phrases, attributes are usually morpho-syntactic properties of the phrases and operations are typical grammatical operations: feature match, movement (becomes), lexical substitution (changes).

3.2. Defining the matrices

Definition 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.

Shape $S(n, l)$ is the shape of the matrix, which consists of n items and each item can be at most of length l.

The length of the items can vary. The items can be sentences or elements in a morphological paradigm. The choice of n depends on how many items need to be shown to illustrate the paradigm and on whether the illustration is exhaustive or sampled.

For example, a matrix of size eight is exhaustive for an agreement problem with three noun phrases and a two-way number differentiation (singular, plural), but can only present a sample of the information for the *spray/load* alternation.

Subject-verb number agreement
Violation of E: wrong subject-verb agreement
Violation of I: wrong agreement on N2 or N3
Violation of R: wrong number of attractors

Spray/load alternation
Violation of E: Wrong lexical choice of preposition
Violation of I: Subject of active voice is not Agent
Violation of R: Wrong number of arguments

Figure 4: Example answer set.

Relational operations

Connective sequential operations, such as alternation or progression are chosen. Their purpose is to transform a list of items (sentences or words) into a predictable sequence that connects all the items.

The values of R, so far, are alternations or progression. They could also be conjunction, disjunction, exclusive OR and other logical or graded operators.

Alternation applies to a given (o, a) pair and loops over all the values of a with a given increment defined over the items of the matrix. For example, the grammatical feature number is binary in certain languages. So, alternation $(o = NP; a_i = (s, p); i = 1, 2, 3, ...)$. This is used to create different alternations of (o, a) in the sentence, which in the subject-verb agreement BLM is used to show independence from linear distance.

Progression applies to countable attributes or ordinal attributes, for example, existence. So, one can have 1,2,...,n of a given object o. Progression can also apply to $position$ or to graded properties such as $length$.

Items The items T are defined by $G_{LP} = (O, A, E, I, L)$ and they are drawn from the set $\mathcal{T}$.

The matrix is created by sampling (o, a, r) triples. The ways in which $r \in R$ can apply to a given (o, a) pair has to be predefined, as it is not entirely context-free.

3.3. Defining the answer set

The answer set W consists of a set of items like those in C. One item in W, w_c, is the correct answer to complete the sequence defined by C. The other items are the contrastive set. They are items that violate G_{LP}, the rules of construction of the context matrix C, either in the primary rules E, in the auxiliary rules I, or in the matrix operators R.

Sometimes they are built almost automatically, sometimes by hand. The cardinality of the answer set is determined by how many facets of the linguistic phenomenon need to be shown to have been learned.

3.4. Augmenting the matrices

Different levels of lexical and structural complexity can be obtained by changing the lexical items (completely or partially), in a given matrix.

Definition An *augmented BLM* is a quadruple (S, R, T, Aug).

S is the shape of the matrix, R are the relational operations that connect the T items of the matrix.

Aug is a set of operations defined to augment the cardinality of $\mathcal{T}$, while keeping S and R constant. Aug is defined by controlled manipulations of Os and As in $\mathcal{T}$ to collect similar elements.

We augment the sentence set $\mathcal{T}$ by modifying the noun phrases of the items in T. We generate alternatives with a language model choosing among the top n, within an acceptability margin from the original sentence. The margin is set with a variable-size window and collects the top 10 alternative noun phrases. The acceptability of the resulting sentences is validated manually. In the next sections, we illustrate the data for two BLM problems, and baseline benchmarking results.

4. Example of two BLM problems

The creation of structured datasets can be a challenging task, depending on the type of linguistic problem being investigated, the available linguistic resources, and the size of the lexical factors involved in the problem. Figure 5 summarizes the pipeline that needs to be followed for the whole process: identifying the data for the linguistic phenomenon under investigation, developing a lexical set seed of lexical items for creating context and answer sets for the BLMS, which are then combined to construct desired context templates and answer sets. From the linguistic phenomenon to the creation of the lexical set seed, various approaches can be pursued based on the type of linguistic phenomenon being investigated. This choice might depend on whether the phenomenon has already been extensively studied in experimental linguistics, the scale of the lexical components involved in the linguistic phenomenon, and the available resources in the target language. We then employ a fill-mask task with transformers to automatically generate additional, plausible constituents for the desired structures.

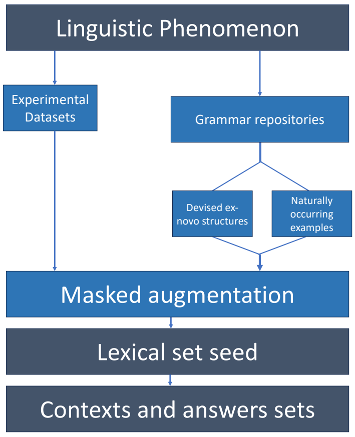

Figure 5: Pipeline for the automatic creation of structured datasets

Figure 9 in the appendix shows an example of the first steps of the process for the BLM-AgrF dataset.

4.1. BLM-AgrF – subject-verb agreement in French

In BLM-AgrF [11], 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 template is illustrated in Figure 2, and an actual example in Figure 6.

The dataset comprises three subsets, of increasing lexical complexity. Type I data is generated based on manually provided seeds (Franck et al. 2002), illustrated in Figure 9, and a template that captures the rules mentioned above. Type II data is generated based on Type I data, by introducing lexical variation with the aid of a transformer, by generating alternatives for masked nouns. Type III data is generated by combining sentences from different instances from the Type II data, while maintaining the structure of the sequence. The structural variations alter the distance and relative depth of the subject and verb and produce a variety of conditions. The different levels of lexical variation will allow us to investigate the impact of lexical variation on the ability of a system to detect grammatical patterns. We include complete instances –

	Context			
1	Il vaso	con il fiore		si è rotto.
2	I vasi	con il fiore		si sono rotti.
3	Il vaso	con i fiori		si è rotto.
4	I vasi	con i fiori		si sono rotti.
5	Il vaso	con il fiore	del giardino	si è rotto.
6	I vasi	con il fiore	del giardino	si sono rotti.
7	Il vaso	con i fiori	del giardino	si è rotto.
8	???			

	Answer set	
1	Il vaso con i fiori e il giardino si è rotto.	coord
2	**I vasi con i fiori del giardino si sono rotti.**	correct
3	Il vaso con il fiore si è rotto.	WNA
4	Il vaso con il fiore del giardino si sono rotti.	AE
5	I vasi con i fiori del giardino si sono rotti.	WN1
6	I vasi con i fiori dei giardini si sono rotti.	WN2

Figure 6: BLM instances for verb-subject agreement, with 2 attractors (*fiore* 'flower', *giardino* 'garden'), with candidate answer set. WNA=wrong number of attractors, AE=agreement error, WN1=wrong nr. for 1st attractor noun (N1), WN2=wrong nr. for 2nd attractor noun (N2)

	Context			
1	NP-Agent	Verb	NP-Theme	PP-Loc
2	NP-Theme	VerbPass	PP-Loc	
3	NP-Agent	Verb	NP-Loc	PP-Theme
4	NP-Loc	VerbPass	PP-Theme	
5	NP-Agent	Verb	NP-Theme	
6	NP-Agent	Verb	PP-Theme	
7	NP-Agent	Verb	NP-Loc	
8	???			

	Answers	
1	**NP-Agent Verb PP-Loc**	correct
2	NP-Agent Verb *PP-Loc	WPrep
3	NP-Agent Verb *[NP-Theme PP-Loc] PP-Loc	WPP
4	NP-Agent Verb NP-Loc because PP-Theme	Adv
5	NP-Agent Verb NP-Loc *PP_{Loc}-Theme	WT
6	*NP-Loc Verb PP-Loc	WS

Figure 7: BLM context template 1 and answers for the spray/load alternation. * = locus of the rule corruption, angled brackets = syntactic embedding. WP= Wrong Preposition; WPP=Wrong Prepositional Phrase; Adv= Adverbial; WT=Wrong Theme; WS=Wrong Subject.

in French – in Appendix A.

4.2. BLM-s/lE.v0 – spray/load verb alternations in English

In the BLM-s/lE problem developed to exhibit the spray/load alternation (discussed in the introduction), each sentence can be described in terms of one distribution-of-three-values rule, governing the semantic roles (Agent, Theme, Locative), and two distribution-of-two-values rules governing syntactic types (nominal phrase NP vs. prepositional phrases PP) and the mood of the verb, whether active (Verb) or passive (VerbPass). We created two templates, targeting the syntax-semantic mapping of the arguments.

In the contrastive answer set, the target sentence is to be chosen from a set of candidates that exhibit minimal differences. The semantic-syntactic mapping of the alternation can be decomposed into a set of smaller patterns that describe the sentences in the alternation and that can be violated to construct incorrect answers. Different subsets of patterns can be used to develop different answer sets.

A variation of this dataset, presented in [12] uses an answer set that change the position of the agent in an active sentence, the type of phrases following the verb, embedding the PP in an NP, and changes in prepositions that introduce different types of arguments.

The dataset presented here, changes subpatterns that govern the correct learning of the syntactic form of the sentence (Wrong Theme, Wrong Subject, Wrong PP in Template 1; SwapLocAgent, NoAgent, SwapThemeAgent, Repeat in Template 2), and others that govern the proper lexical selection (Wrong Preposition, Adverbial).

Different answer sets, that focus on different rule subpatterns, will allow for detailed investigations of the type of information that is more easily or more difficult to detect, and to determine principles of designing the answer set for the most informative phenomenon investigation.

Like the BLM-AgrF dataset, the BLM-s/lEv.0 dataset presents lexically varied versions, Type I, II, and III, of increasing variability. The structure of the context and answer set of one alternant is presented in Figure 7. Figure 11 for the other template and relevant lexical examples of both templates are given in appendix B.

5. Benchmarking systems

Our goal is to investigate – and ultimately use this knowledge to enhance – textual representations built using pretrained large language models. To determine whether such representations encode linguistic rules, and to what degree they are compositional, we use BLM tasks that provide data generated using specific rules, and baseline systems that should be capable of detecting the patterns that encode the relevant rules for the targeted phenomena in the distributed continuous sentence representations. We choose a FFNN and a CNN as our baselines. The FFNN should be able to discover patterns distributed throughout a sentence, and throughout a sequence of sentences, while the CNN could discover localized patterns – both in the sentence and the sequence.

As presented above, a BLM problem instance consists of a context and an answer set. The context is a sequence of 7 sentences, and the answer set is a set of 6 sentences, one of which is a correct continuation of the input sequence. All sentences are encoded using BERT [13] – as the embedding of the [CLS] token on the last layer of the model. We used the pretrained "BERT-base-multilingual-cased" model[1]. The sentence representations are combined in different ways, depending on the baseline system – a FFNN or a CNN – used.

The input to the FFNN is the concatenation of sentence embeddings in the BLM instance context, as a vector of size 7 * 768. This input is processed through 3 fully connected layers, which progressively compress the input size (7 * 768 $\xrightarrow{layer1}$ 3.5 * 768 $\xrightarrow{layer2}$ 3.5 * 768 $\xrightarrow{layer3}$ 768) to obtain the size of a sentence representation. The FFNN's interconnected layers enable it to capture patterns that are distributed throughout the entire input vector.

The input to the CNN is the stacked sentence embeddings in the BLM instance context, as a (7 x 768) array. This input undergoes three consecutive layers of 2-dimensional convolutions, where each convolutional layer uses a kernel size of (3x3) and a stride of 1, without dilation. The resulting output from the convolutional process is then passed through a fully connected layer, which compresses it to the size of the sentence representation (768). By using a kernel size of (3x3), stride=1, and no dilation, this configuration emphasizes the detection of localized patterns within the sentence sequence array.

The output of both systems is a vector representing a sentence embedding. This is compared to the sentence representations in the answer set, and the one with the highest score is considered the correct answer. Details are included in appendix C.

6. Results

Previous published work from our group and current ongoing work has benchmarked the problems generated by these datasets and analysed the errors, with interesting results [11, 14].

We report here the novel results on the BLM-s/lE dataset, which are qualitatively similar to those reported for BLM-AgrF [11], thus confirming some general trends. Figure 8 shows the results. The top panels shows results using all the data, the bottom panel shows results using only data sizes that match type I data size, hence smaller data sizes for type II and type III.

Globally, the results are very good. But interesting differences emerge if we vary data sizes. If we train on all data, more lexically-varied data (types II and III) give better results, but if we train on equally sized datasets, we see an improvement of the results in training on type II data whether testing on type II or III. It appears, then, that on smaller datasets, the template patterns is perhaps better learnt in type II, while still retaining the notion of lexical variation that is lacking in type I.

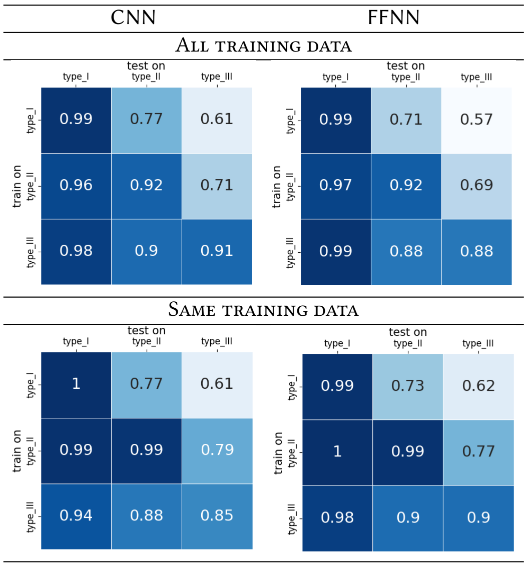

Figure 8: F1 results (averages over 5 runs) for alternant one.

Inspecting the increase in performance with different training data sizes, shown in Figure 12 in the appendix, it is confirmed that learning is very fast and plateaus already with a few thousand examples for all train-test combinations, with the exception of training on Type I and testing on type III, which is clearly too difficult.

Both baseline systems lead to good results despite the variations in the input – structural and lexical – that superficially obfuscate these phenomena, and the near-miss incorrect answers, confirming that the phenomena we target are encoded in the sentence representations. Because of their different architectures and the type of patterns they discover, the high performance of both systems indicates that relevant patterns for our two targeted phenomena are localized in BERT sentence embeddings. Further steps can take advantage of the structured way the data was constructed to attempt to disentangle the various generative rules and additional factors in the inputs.

7. Related Work

Previous work has focussed on understanding the automatic learning of verb alternations in terms of syntactic and semantic properties of the verbs and their argument structures [15]. These properties have been explored in

[1] https://huggingface.co/bert-base-multilingual-cased

relation to their representation in LLMs, across various dimensions of performance for different models [16, 17]. In particular, [17] suggest that LLMs with contextual embeddings encode linguistic information on verb alternation classes, at both the word and sentence levels. In their work, [17] build upon [16] observations and highlight the superior performance of one transformer Electra [18] compared to other large language models.

The automatic generation of RPM-like matrices, whether in vision or in language, is technically challenging. In computer vision, several formalisms have been proposed ([19] formulate RPMs with first-order logic; [20] propose Procedurally Generated Matrices (PGM) datasets through relation-object-attribute triple instantiations; [21] use the Attributed Stochastic Image Grammar (A-SIG [22]). Structured synthetic datasets have been mostly developed to study issues of generalisation and disentaglement, in vision [23], with full-fledged experimentation and for language in a preliminary, nonRPM -like dataset, consisting of simple examples containing a few morphological markings [24]. The simplicity of the sentences does not provide a sufficiently realistic challenge from a linguistic point of view. Very recent work has started exploring the picture-naming potential of language to solve problems in vision [25].

8. Conclusions

In this paper, we have presented the new BLM task, provided its formal specifications and illustrated the first instances of BLM problems and benchmarking results with baseline architectures. Current work is developing new dedicated architectures based on Variation Autoencoders [14] and developing new BLM problems. Future work lies in further automating the data development pipeline, to make the creation on BLM data sets also accessible to less computationally-oriented linguists and investigating the structure and nature of the information encoded in the learned inner representations.

Acknowledgments

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

[1] Y. Lakretz, G. Kruszewski, T. Desbordes, D. Hupkes, S. Dehaene, M. Baroni, The emergence of number and syntax units in LSTM language models, arXiv preprint arXiv:1903.07435 (2019).

[2] Y. Lakretz, D. Hupkes, A. Vergallito, M. Marelli, M. Baroni, S. Dehaene, Mechanisms for handling nested dependencies in neural-network language models and humans., Cognition (2021). doi:2021 10.1016/j.cognition.2021.104699.

[3] M. Sablé-Meyer, J. Fagot, S. Caparos, T. van Kerkoerle, M. Amalric, S. Dehaene, Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity, Proceedings of the National Academy of Sciences 118 (2021). doi:10.1073/pnas.2023123118.

[4] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.

[5] B. Levin, English verb classes and alternations: A preliminary investigation, University of Chicago Press, 1993.

[6] J. Beavers, The spray/load alternation, The Wiley Blackwell Companion to Syntax, Second Edition (2017) 1–31.

[7] P. Merlo, A. An, M. A. Rodriguez, Blackbird's language matrices (BLMs): a new benchmark to investigate disentangled generalisation in neural networks, ArXiv: cs.CL 2205.10866 (2022). URL: https://arxiv.org/abs/2205.10866. doi:10.48550/ARXIV.2205.10866.

[8] P. Merlo, Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Motivations and formal specifications, ArXiv cs.CL 2306.11444 (2023). URL: https://doi.org/10.48550/arXiv.2306.11444. doi:10.48550/arXiv.2306.11444.

[9] P. Merlo, C. Jiang, G. Samo, V. Nastase, Blackbird Language Matrices Tasks for Generalization, in: GenBench: The first workshop on (benchmarking) generalisation in NLP, Singapore, 2023.

[10] J. C. Raven, Standardization of progressive matrices, British Journal of Medical Psychology 19 (1938) 137–150.

[11] A. An, C. Jiang, M. A. Rodriguez, V. Nastase, P. Merlo, 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, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1363–1374. URL: https://aclanthology.org/2023.eacl-main.99.

[12] G. Samo, V. Nastase, C. Jiang, P. Merlo, 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, Singa-

pore, 2023.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] V. Nastase, P. Merlo, Grammatical information in BERT sentence embeddings as two-dimensional arrays, in: Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023), Toronto, Canada, 2023.

[15] O. Majewska, A. Korhonen, Verb classification across languages, Annual Review of Linguistics 9 (2023) 313–333. doi:10.1146/annurev-linguistics-030521-043632.

[16] K. Kann, A. Warstadt, A. Williams, S. R. Bowman, Verb argument structure alternations in word and sentence embeddings, in: Proceedings of the Society for Computation in Linguistics (SCiL) 2019, 2019, pp. 287–297. URL: https://aclanthology.org/W19-0129. doi:10.7275/q5js-4y86.

[17] D. Yi, J. Bruno, J. Han, P. Zukerman, S. Steinert-Threlkeld, 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, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 142–152. URL: https://aclanthology.org/2022.blackboxnlp-1.12.

[18] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre- training text encoders as discriminators rather than generators, in: ICLR, 2020, pp. 1–18.

[19] K. Wang, Z. Su, Automatic generation of raven's progressive matrices, in: Q. Yang, M. J. Wooldridge (Eds.), Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, AAAI Press, 2015, pp. 903–909. URL: http://ijcai.org/Abstract/15/132.

[20] D. Barrett, F. Hill, A. Santoro, A. Morcos, T. Lillicrap, Measuring abstract reasoning in neural networks, in: J. G. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 4477–4486. URL: http://proceedings.mlr.press/v80/santoro18a.html.

[21] C. Zhang, F. Gao, B. Jia, Y. Zhu, S. Zhu, RAVEN: A dataset for relational and analogical visual reasoning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 5317–5327. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Zhang_RAVEN_A_Dataset_for_Relational_and_Analogical_Visual_REasoNing_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00546.

[22] S. Zhu, D. Mumford, A stochastic grammar of images, Found. Trends Comput. Graph. Vis. 2 (2006) 259–362. URL: https://doi.org/10.1561/0600000018. doi:10.1561/0600000018.

[23] S. van Steenkiste, F. Locatello, J. Schmidhuber, O. Bachem, Are disentangled representations helpful for abstract visual reasoning?, in: NeurIPS 2019, 2020.

[24] A. M'Charrak, Deep Learning for Natural Language Processing (NLP) using Variational Autoencoders (VAE), Master's thesis, ETH Switzerland, 2018. URL: https://pub.tik.ee.ethz.ch/students/2018-FS/MA-2018-22.pdf.

[25] X. Hu, S. Storks, R. Lewis, J. Chai, In-context analogical reasoning with pre-trained language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1953–1969. URL: https://aclanthology.org/2023.acl-long.109.

[26] J. Franck, G. Vigliocco, J. Nicol, Subject-verb agreement errors in french and english: The role of syntactic hierarchy, Language and cognitive processes 17 (2002) 371–404.

A. BLM-AgrF problem

Example subject NPs from [26]
L'ordinateur avec le programme de l'experience The computer with the program of the experiments

Manually expanded and completed sentences
L'ordinateur avec le programme de l'experience est en panne. The computer with the program of the experiments is down.
Jean suppose que l'ordinateur avec le programme de l'experience est en panne. Jean thinks that the computer with the program of the experiments is down.
L'ordinateur avec le programme dont Jean se servait est en panne. The computer with the program that John was using is down.

A seed for language matrix generation				
Jean suppose que Jean thinks that	*l'ordinateur* the computer	*avec le programme* with the program	*de l'experience* of the experiment	*est en panne* is down
	les ordinateurs the computers	*avec les programmes* with the programs		*sont en panne* are down

Figure 9: Examples from [26], manually completed and expanded sentences based on these examples, and seeds made based on these sentences for the subject-verb agreement BLM-AgrF dataset that contain all number variations for the nouns and the verb.

Contexts	
Example	Translation
1 La conférence sur l'histoire a commencé plus tard que prévu.	*The talk on history has started later than expected.*
2 Les responsables du droit vont démissionner.	*Those responsible for the right will resign.*
3 L' exposition avec les peintures a rencontré un grand succès.	*The show with the paintings has met with great success.*
4 Les menaces de les réformes inquiètent les médecins.	*The threats of reforms worry the doctors.*
5 Le trousseau avec la clé de la cellule repose sur l'étagère.	*The bunch of keys of the cell sits on the shelf.*
6 Les études sur l'effet de la drogue apparaîtront bientôt.	*The studies on the effect of the drug will appear soon.*
7 La menace des réformes dans l' école inquiète les médecins.	*The threat of reforms in the school worries the doctors.*
Answers	
Example	Translation
1 Les nappes sur les tables et le banquet brillent au soleil.	*The tablecloths on the table and the console shine in the sun.*
2 **Les copines des propriétaires de la villa dormaient sur la plage.**	*The friends of the owners of the villa were sleeping on the beach.*
3 Les avocats des assassins vont revenir.	*The laywers of the murderers will come back.*
4 Les avocats des assassins du village va revenir.	*The lawyers of the murderers of the village will come back.*
5 La visite aux palais de l' artisanat approchent.	*The visit of the palace of the crafts is approaching.*
6 Les ordinateurs avec le programme des expériences sont en panne.	*The computers with the program of the experiments are broken.*

Figure 10: Example of lexically varied contexts for the main clause contexts for the subject-verb agreement BLM-AgrF dataset. Correct answer in bold.

B. BLM-s/lE

Context			
1 NP-Agent	Verb	NP-Theme	
2 NP-Agent	Verb	PP-Theme	
3 NP-Agent	Verb	NP-Loc	
4 NP-Agent	Verb	PP-Loc	
5 NP-Agent	Verb	NP-Theme PP-Loc	
6 NP-Theme	VerbPass	PP-Loc	
7 NP-Agent	Verb	NP-Loc	PP-Theme
8 ???			

Answers	
1 **NP-Loc VerbPass PP-Theme**	correct
2 NP-Loc Verb *NP-Agent PP-Theme	SLA
3 NP-Loc VerbPass *PP$_{Loc}$-Theme PP-Loc	WP
4 NP-Loc VerbPass *PP-Loc	Repeat
5 *NP-Theme Verb NP-Loc	NoAgent
6 NP-Theme Verb *NP-Agent PP-Loc	STA

Figure 11: BLM context template 2 and answers for the spray/load alternation. * = locus of the rule corruption, angled brackets = syntactic embedding. WP=WrongPrep, SLA=SwapLocAgent, STA=SwapThemeAgent.

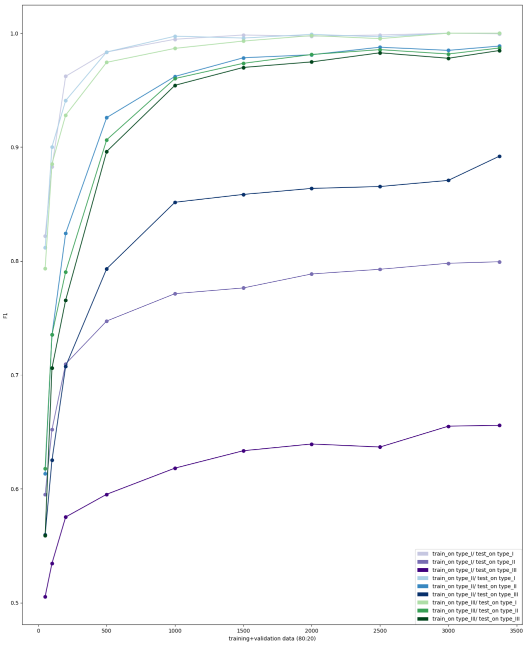

Figure 12: Effect of training data size on the BLM-S/lE datasets.

Template 1, Type I

Context

The crew sprayed some water into a plastic container.
Some water was sprayed into a plastic container.
The crew sprayed a plastic container with some water.
A plastic container was sprayed with some water.
The crew sprayed some water.
The crew sprayed with some water.
The crew sprayed a plastic container.
???

Answers

The crew sprayed into a plastic container.
The crew sprayed under a plastic container.
The crew sprayed some water from rivers into a plastic container.
The crew sprayed a plastic container because of some water.
The crew sprayed a plastic container under some water.
A plastic container sprayed into a plastic container.

Figure 13: Template 1 -Type I context and answer set.

Template 1, Type II

Context

Katrina sprayed some liquid into the windshield.
Some of the chemicals were sprayed into the windshield.
I sprayed the wall with some water.
The windshield was sprayed with some water.
The artist sprayed some of the material.
The artist sprayed with some of the chemicals.
Someone sprayed the sink.
???

Answers

The artist sprayed onto the wall.
The crew sprayed under the bathroom.
Katrina sprayed some of the chemicals for the refinery into the windshield.
Someone sprayed the bathroom because of some water.
Katrina sprayed a plastic container under some liquid.
The bathroom sprayed into the bathroom.

Figure 14: Template 1 -Type II context and answer set.

Template 1, Type III

Context

The crew sprayed some water into a plastic container.
Heavier material was sewed onto the straps.
Water spurts the room with dirt.
It was pumped with steam.
Archaeologists clean the filter monthly and swash some tank water.
Scientists promote a strategy to seed with sulfur.
The volunteers swash the biomedia.
???

Answers

The Egyptians sow on the soil.
Archaeologists clean the filter monthly and swash under the sink.
The crew sprayed some of the paint from vegetables into the windshield.
The man smeared the walls because of the flour.
The waitress sprinkles eggs under sugar.
The windows strung over the windows.

Figure 15: Template 1 -Type III context and answer set.

C. Architectural Specifications

All systems used a learning rate of 0.001 and Adam optimizer, and batch size 100. The training was done for 120 epochs. The experiments were run on an HP PAIR Workstation Z4 G4 MT, with435 an Intel Xeon W-2255 processor, 64G RAM, and a MSI GeForce RTX 3090 VENTUS 3X OC 24G GDDR6X GPU.

We tested BERT sentence embeddings with baseline CNN and FFNN baseline architectures. [13]. The sentence embeddings are the encoding of the [CLS] token on the last layer of the model.

The FFNN receives the input as a concatenation of sentence embeddings in a sequence, with a size of 7 * 768. This input is then processed through 3 fully connected layers, which progressively compress the input size (7 * 768 $\rightarrow layer1$ 3.5 * 768 $\rightarrow layer2$ 3.5 * 768 $\rightarrow layer3$ 768) to obtain the size of a sentence representation. The FFNN's interconnected layers enable it to capture patterns that are distributed throughout the entire input vector.

The CNN takes as input an array of embeddings with a size of (7 x 768). This input undergoes three consecutive layers of 2-dimensional convolutions, where each convolutional layer uses a kernel size of (3x3) and a stride of 1, without dilation. The resulting output from the convolutional process is then passed through a fully connected layer, which compresses it to the size of the sentence representation (768). By using a kernel size of (3x3), stride=1, and no dilation, this configuration emphasizes the detection of localized patterns within the sentence sequence array.

Both networks produce the same output, which is a vector representing the sentence embedding of the correct answer. The objective of learning is to maximize the probability of selecting the correct answer from a set of candidate answers. To achieve this, we employ the max-margin loss function, considering that the incorrect answers in the answer set are intentionally designed to have minimal differences from the correct answer. This loss function combines the distances between the predicted answer and both the correct and incorrect answers. Initially, we calculate a score for each candidate answer's embedding e_i in the answer set $\mathcal{A}$ with respect to the predicted sentence embedding e_{pred}. This score is determined by the cosine of the angle between the respective vectors:

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

The loss function incorporates the max-margin concept, taking into account the difference between the score of the correct answer e_c and each of the incorrect answers e_i:

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

During prediction, the answer with the highest score value from the candidate set is selected as the correct answer.