Multilingual vs. monolingual transformer models in encoding linguistic structure and lexical abstraction

Vivi Nastase[1,*], Giuseppe Samo[1], Chunyang Jiang[1,2] and Paola Merlo[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]University of Geneva, Geneva, Switzerland

Abstract

Multilingual language models are attractive, as they allow us to cross linguistic boundaries, and solve tasks in different languages in the same mathematical space. They come, however, at a cost: in the quest to find a shared space that satisfies (to a certain degree) all languages, the resulting representations lose, or fail to capture, properties specific to each language. We present an investigation into detecting linguistic structure through lexical abstraction. We study both a multilingual and a monolingual language model, and quantify the loss of information between them.

I modelli di linguaggio multilingue permettono di oltrepassare i confini linguistici e di risolvere task in lingue diverse mantenendo lo stesso spazio matematico. Tuttavia, questi modelli hanno un costo: nella ricerca di uno spazio condiviso che soddisfi (in una certa misura) tutte le lingue, le rappresentazioni risultanti perdono, o non riescono a catturare, le proprietà specifiche di ciascuna lingua. Usando il fenomeno di astrazione lessicale, presentiamo qui un'indagine su come la struttura linguistica venga individuata: analizziamo sia un modello linguistico multilingue che un modello monolingue, e quantifichiamo la perdita di informazioni tra di essi.

Keywords

multilingual and monolingual models, linguistic abstraction, functional words

1. Introduction

Multilingual models are attractive because they project all languages represented in the training data into the same n-dimensional space. This makes it easy to plug them into tasks in different languages.

The abilities of multilingual models are being actively debated. The first large-scale multilingual models suffered from *the curse of multilinguality*: "more languages leads to better cross-lingual performance on low-resource languages up until a point, after which the overall performance on monolingual and cross-lingual benchmarks degrade" [1, p. 1], which could be remedied by increasing the capacity of the models [1], or by training bilingual models for low-resource languages, where each such language is paired with a linguistically-related language [2]. Forcing many languages to share the parameter space, may lead to the emergence of language universal representations in pretrained encoder models [3], possibly even grammatical structure [4, 5]. However, these models do not encode structure in a language-independent, abstract, way, but rather encode language-specific token-level clues [6].

The work presented in this paper adds more detail to this picture. We investigate how accessible sentence structure is in sentence representations, comparing the representations obtained from a multilingual encoder model to its monolingual counterpart. We conduct this exploration on the problem of *lexical abstraction*, the process of reducing a sentence to its syntactic and semantic "skeleton" by replacing noun and prepositional phrases with functional words, as in the example: *The authors wrote the paper.* and *They wrote it.* We expect that lexical abstraction has occurred if we can detect the same syntactic structure in the embeddings of lexicalized and functional versions of pairs of sentences. This setup verifies whether the multilingual model or the monolingual models perform better. The former results would indicate that training on several languages is beneficial to discovering shared structures. The latter result, instead, would indicate that sentence structure is encoded in a more language-specific manner, and is encoded better by a monolingual model, as the model does not need to reconcile the different ways the same type of grammatical information is expressed in different languages (e.g. number, case, gender, definiteness).

To further explore multilingual models, we also perform experiments with generative LLMs, as they have been shown to favour English as an "internal" language [7, 8]. Here, we test whether a multilingual LLM detects (and generates) sentence structure better in English sentences than Italian ones, by prompting the model with English, and separately with Italian sentences, asking it to produce the Italian functional form.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
vivi.a.nastase@gmail.com (V. Nastase); giuseppe.samo@idiap.ch (G. Samo); chunyang.jiang42@gmail.com (C. Jiang); Paola.Merlo@unige.ch (P. Merlo)

2. Data

To investigate how accessible sentence structure is in representations built by large language models, we use the Italian portion of a dataset that models the verb alternations change-of-state (CoS) and object drop (OD) [9]. The CoS verb class can undergo the transitive/intransitive causative alternation, where the object of the transitive verb bears the same semantic role (Patient) as the subject of the intransitive verb (*The tourist broke the vase*/*The vase broke*). The transitive form of the verb has a causative meaning. In contrast, for OD verbs the subject bears the same semantic role (Agent) in both the transitive and intransitive forms and the verb does not have a causative meaning (*The artist was paiting this fresco/ The artist was painting*) [10, 11]. Italian shows the same asymmetry but marks the intransitive alternant for CoS with a reflexive-like element SI (*Il turista ruppe il vaso/Il vaso si ruppe*; *L'artista stava dipingendo questo affresco/l'artista stava dipingendo*).

These verb classes constitute an ideal test-bed for our research question, because their combination of syntactic and semantic structure allows us not only to test whether sentences with different syntactic structures can be distinguished, but also whether sentences with the same syntactic structure but differing in the semantic roles can be distinguished.

The data, described in detail in [12], consists of instances of a Blackbird Language Matrices (BLM), a linguistic puzzle [13]. Each instance consists of an input *context* of seven sentences that illustrate several variations of CoS/OD verbs, and an *answer set* that contains a correct answer, and nine wrong answer candidates, each of which is erroneous in specific ways. Figure 1 shows the syntactic-semantic structure of the sentences in a BLM instance. Lexicalized and functional instances are shown in tables 4 and 5 in the appendix.

Each BLM instance has a lexicalized (LEX) and a functional (FUN) form. In addition, there are three variations – type I, type II, type III – with increasing levels of lexical variation. The dataset is built based on thirty (manually chosen) verbs from each of the two classes discussed in Levin [10]. The functional lexicon has been manually selected by the authors to maintain the syntactic and semantic acceptability of the sentences.

We build two variations starting from this dataset that allow us to test, from several angles, whether sentence structure is encoded in a sentence embedding in an abstract manner.

Sentences We compile parallel versions of the sentences in their lexicalized and functional word forms from the FUN and LEX subsets of the type I BLM dataset. Each sentence has associated its syntactic pattern (the syntactic version of the syntactic-semantic template shown in Figure 1). From these, we sample 6000 sentences, uniformly distributed over the eight syntactic-semantic patterns. These are split into 4800:1200 training and test instances and 20% of the training data is used for validation (train:dev:test – 3840:960:1200).

COS context			
1 Agent	Active	Patient	p-NP
2 Agent	Active	Patient	by-NP
3 Patient	Passive	by-Agent	p-NP
4 Patient	Passive	by-Agent	by-NP
5 Patient	Passive		p-NP
6 Patient	Passive		by-NP
7 Patient	Active		p-NP
8 ?			

COS answers				
Patient	SI Active	by-NP	**Correct**
Agent	SI Active	by-NP	I-Int
Patient	Passive	by-Agent	ER-Pass
Agent	Passive	by-Patient	IER-Pass
Patient	Active	Agent	R-Trans
Agent	Active	Patient	IR-Trans
Patient	Active	by-Agent	E-WrBy
Agent	Active	by-Patient	IE-WrBy
Patient	Active	by-NP	NoSI
Agent	Active	by-NP	I-NoSI

OD context			
1 Agent	Active	Patient	p-NP
2 Agent	Active	Patient	by-NP
3 Patient	Passive	by-Agent	p-NP
4 Patient	Passive	by-Agent	by-NP
5 Patient	Passive		p-NP
6 Patient	Passive		by-NP
7 Agent	Active		p-NP
8 ?			

OD answers			
Patient	Active	by-NP	I-Int
Agent	Active	by-NP	**Correct**
Patient	Passive	by-Agent	IER-Pass
Agent	Passive	by-Patient	ER-Pass
Patient	Active	Agent	I-Trans
Agent	Active	Patient	R-Trans
Patient	Active	by-Agent	IE-WrBy
Agent	Active	by-Patient	E-WrBy
Patient	SI Active	by-NP	I-SI
Agent	SI Active	by-NP	SI

Figure 1: Context and answer sentence structures for change-of-state (CoS) verbs (left), and object drop (OD) verbs (right).

BLM data Of the thirty verbs for each class, change of state and object drop, three are selected for testing and the other 27 for training. All instances for the three testing verbs are used. Two-thousand instances of the other 27 verbs are randomly sampled for training. Ten percent of the training data is dynamically selected for validation. The same 27:3 verb split is used for all FUN/LEX and type I/type II/type III variations. All variations have 2000 instances for training, 300 for testing. In the experiments reported here we use a variation where the CoS and OD subtasks are merged. The data is split in a similar manner for training and testing (and using the same verbs for training and testing as in the split of the individual subsets).

3. Experiments

We aim to quantify to what degree multilingual and monolingual language models encode syntactic structure by using the lexical abstraction property of pronouns and adverbs relative to nouns and noun phrases. We explore encoder models, and test whether the same syntactic structure and semantic role information is encoded in the embeddings of lexicalized sentences and their functional versions. With generative LLMs, we compare the performance of a model in generating the functional version of an input sentence, when this input is either in English or Italian, and the output is constrained to be Italian.

3.1. Sentence structure in encoder models

We perform two analyses to test whether the representation of functional and lexicalized sentences encode the same grammatical structure, in the same way: (i) we analyze individual sentences and test to what degree their grammatical structure (phrases and their semantic roles) can be detected (Section 3.1.1); (ii) we deploy the BLM linguistic puzzles, whose solution relies on detecting shared structure at the level of input sequence and within each sentence (Section 3.1.2).

We obtain word and sentence representations (as averaged token embeddings) from an Electra pretrained model [14][1]. We choose Electra because it has been shown to perform better than models from the BERT family on the Holmes benchmark[2], and to also encode information about syntactic and argument structure better [15, 16]. We use the Italian Electra[3] as our monolingual model.

3.1.1. Grammatical structure in sentence embeddings

Syntactic structure and semantic roles represent complex information, which may be encoded by weighted combinations of subsets of dimensions [17, 18].

We mine the sentence repesentations for this information following the approach described in Nastase and Merlo [16]. Using a variational encoder-decoder, an input sentence is compressed into a representation that captures syntactic and semantic role information, by imposing that the system reconstructs a sentence with the same syntactic and semantic information. An instance consists of an input sentence s_i with structure str_i, and a set of candidate outputs, with a sentence $s_j \neq s_i$ that has the same structure ($str_j = str_i$), and N negative examples s_k that have different structures ($str_k \neq str_i$). In our experiments we use N = 7. The structure information is used to build the dataset and obtain a deeper evaluation of the results, but is not provided to the system.

Using the sentences datasets described in section 2, we built datasets consisting of a mix of FUN and LEX instances (an instace will only contain either FUN or LEX sentences), and use the above-mentioned set-up to test: (i) how well a system reconstructs a sentence with the desired syntactic and semantic information, measured at the output through F1 score[4], and (ii) how well the system identifies the different patterns. Specifically, we ask whether the same patterns in lexicalized and functional forms are detected as being the same, and, thus, mapped onto the same representation on the latent layer. We estimate similarity of representations by visualising them on the latent layer. Sentence embeddings from Electra have size 768, and the latent layer in the used system has size five.

test on / train on	FUN		LEX	
	e	e-It	e	e-it
FUN	0.92	0.98	0.20	0.23
LEX	0.20	0.32	0.78	0.92
Mixed	0.76	0.91	0.57	0.81

Table 1
F1 scores (averages over three runs) on predicting the sentence with the same structure as the input, through a variational encoder-decoder system, for sentences encoded with (multilingual) Electra (e) or (monolingual) Electra-It (e-It).

Table 1 shows the averaged F1 scores over three experiments. We note first that training and testing on the same type (FUN or LEX) leads to high results, thus validating the experimental set-up.

The results on test data of the same type as the training are very different from those on the test of the other type. This indicates that for each of the FUN and LEX data variations, the system discovers different clues to match two sentences with the same structure. The high results when training on the sentences with functional words may also indicate overfitting because of the repetitive vocabulary. We note that, consistently, the results obtained when using a monolingual model are higher than those when using the multilingual one, despite the assumption that a multilingual model must learn more abstract representations to satisfy the constraints of modeling many languages.

Additional information comes from the analysis of the compressed representations on the latent layer, which are expected to capture the sentence structure that is shared by the functional and lexicalized data. We show the projection on the latent layer of the sentence representations in Figure 2, when sentence representations are obtained from Electra (left) and Electra-It (right). We note that these latent projections cluster by the syntactic structure and semantic roles of the sentences, and that using Electra-It representations leads to a tighter mix of lexicalized and functional sentences that have the same syntactic structure. This adds depth to the results in Table 1 – showing that when trained on a mix of functionalized and lexicalized instances, the system is able to discover a shared space of clues about the grammatical structure – and also shows that in the representations obtained from Electra-It there are stronger shared clues about grammatical structure in both functionalized and lexicalized sentences compared to the multilingual Electra model.

[1] *google/electra-base-discriminator*
[2] The HOLMES benchmark leaderboard: https://holmes-leaderboard.streamlit.app/. At the time of writing, the ranks were: Electra - 16, DeBERTa - 21, BERT - 41, RoBERTa - 45.
[3] *dbmdz/electra-base-italian-xxl-cased-discriminator*
[4] When processing each instance, the system chooses among 8 options, of which one is correct. The F1 score of the "positive" class provides the most balanced measure of performance.

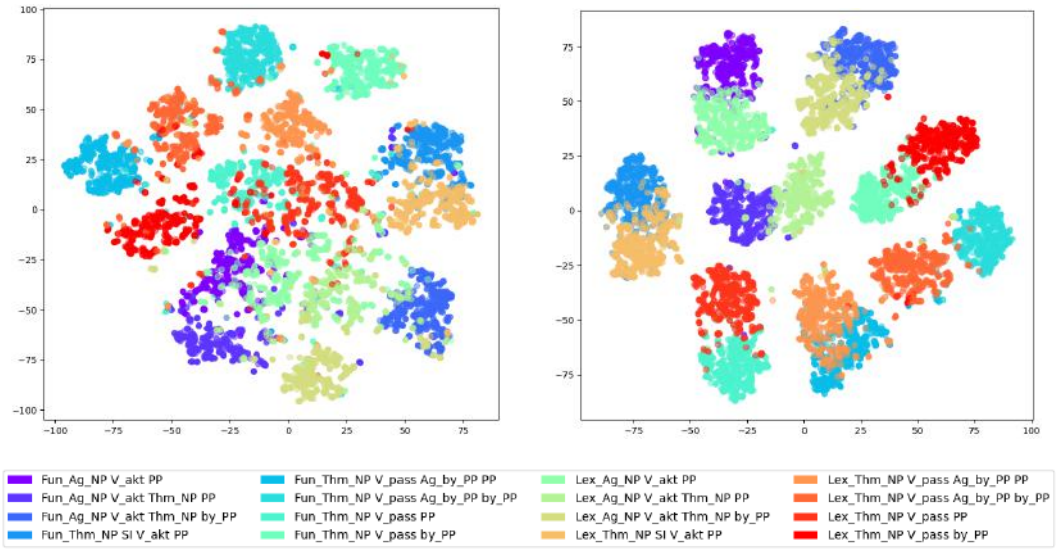

Figure 2: Latent representation analysis: t-SNE projection of vectors on the latent layer for the sentences in the training instances, when sentences are encoded using electra (left) vs. electra-it (right). Lexicalized (Lex) and functional (Fun) sentences with the same syntactic-semantic pattern should ideally be projected onto close vectors in the latent space.

3.1.2. Task solving

It might be objected that the previous experiments and visualisations do not conclusively show that latent representations encode structure, as opposed to just distinguishing seven distinct but amorphous classes. We use the BLM data to provide additional support to the conclusion that structure is represented. The BLM task frames a linguistic phenomenon as a linguistic puzzle. Solving this puzzle relies on detecting the linguistic objects, their relevant properties, and the structure both within each sentence, and across the input sequence.

Our BLM dataset has several levels of complexity: (i) a mixture of change-of-state and object-drop verbs, which exhibit different semantic frames for the intransitive answers (patient vs agent subjects), and share other frames (see Figure 1); (ii) lexicalized and functional instances; (iii) maximal level of lexical variation in each instance. This set-up will allow us to test whether syntactic structure and semantic roles are encoded similarly in the representation of lexicalized and functional sentences by monolingual and multilingual encoder models.

We use the system described by Nastase and Merlo [16], that solves the BLM problem in two steps: compresses the sentence into a representation that encodes the structure relevant to the BLM puzzle – linguistic objects and their syntactic and semantic role properties –, and uses these compressed representations to solve the multiple-choice puzzle. The system's two steps are encoded through interconnected variational encoder-decoders, as illustrated in Figure 4, which are trained together. The learning

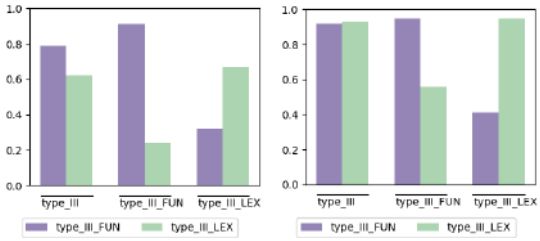

Figure 3: Comparison between the multilingual (left) and monolingual (right) electra models for solving the BLM task: average F1 over three runs. x-axis shows the traininng data: training on FUN and LEX instances jointly vs. training separately on FUN and LEX

objective is to maximize the score of the correct answer from the candidate answer set, and minimize that of the incorrect ones. During testing, the system constructs the representation of an answer, then chooses the closest one from the given options. All potential answers consist of a verb frame filled with phrases that play specific roles (Section 2). The correct one consists of the combination of phrases whose roles fit together for the given verb, while the other contain similar pieces, but which violate some semantic, syntactic (or both) rules. This set-up allows us to test whether specific elements in the sentences from the input sequence, and their semantic roles have been detected and used properly in building the correct answer.

Figure 3 shows the F1 results (as averages over three

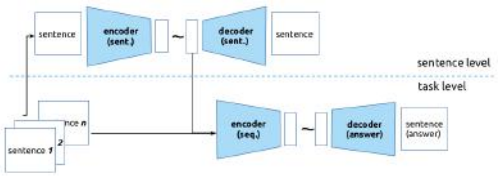

Figure 4: Two-step VAE BLM solver

runs) of training jointly on FUN and LEX instances vs. separate training for the causal verbs BLM task. We use the dataset variations that have the maximum lexical variation (type III, see Section 2), to encourage the system towards finding more abstract representations.

Processing separately datasets of sentences with and without functional words leads to high results within each task, validating the experimental set-up, but leads to low results when testing across tasks. This shows, as in the analysis of the sentences datasets, that for each of the FUN and LEX subsets, the systems discovers and exploits different regularities in the training data, despite the high degree of lexical variation in the lexicalized subset. Using a mixed training dataset, instead, encourages all systems to find a shared feature space. As in the experiments on finding structure in the individual sentences, we note that the shared structure between functional and lexicalized sentences is better encoded in the monolingual Electra model, compared to the multilingual version. Furthermore, comparing the results on the separate training (FUN vs. LEX), we note that the monolingual representations lead to much better generalizations for both set-ups, as the model trained only on the functional forms leads to a significant performance increase when applied on the lexicalized data: from 0.24 average F1 to 0.56 on the monolingual model. The system is also much better able to generalize when trained on the lexicalized version only with the monolingual model: 0.95 average F1 score vs. 0.67 for the multilingual model.

3.2. Generating functional variations of sentences

Multilingual generative models are not exposed to the same amounts of training data across languages and probably for that reason they do not appear to treat every language in their training data equally. In fact, evidence has shown that English serves as a latent language for generative models (LlaMa 2). Tracking an input in languages other than English through the intermediate layers of the transformer, it has been shown that from the input the representations drift more and more towards English, with a switch towards the input language's representation only at the last layers [7, 8]. We test whether this implies that the structure of an English sentence is encoded better than the structure of a sentence in Italian, or whether they both benefit from having been encoded together. For this we prompt the models with lexicalized sentences, and instruct them to convert the sentences to their functional equivalents by replacing nouns with pronouns, prepositional phrases with adverbs and deictics, while maintaining the syntactic structure.

From the dataset of sentences described in section 2, we build 110 instances, each consisting of an Italian sentence, its English translation, and the corresponding Italian functional form. We use 100 instances for testing, and from the remaining 10 we sample N for N-shot prompting (N $\in \{1, 5\}$).

3.2.1. Prompts

We use Meta-Llama-3.1-8B-Instruct, trained on diverse multilingual data with general instruction-following capabilities, and compare two settings: (i) prompting in English with English sentences and requesting Italian functional forms, (ii) prompting in Italian, with Italian sentences, and requesting the corresponding Italian functional form. We use batch processing with fixed batch sizes of five to ensure consistent evaluation conditions across all experiments.

The prompt with English input sentence, and requesting an Italian functional version is shown below.

> Convert these English sentences to Italian by replacing noun phrases with pronouns and prepositional phrases with adverbs. Keep the same syntactic structure.
>
> Examples:
>
> Input: "these toys were carved by his parents in the cabin" → Output: "questi erano intagliati da loro là"
>
> Now convert these:
>
> 1. Input: "that song had been hummed by my friends for a few weeks"
> Output:
>
> 2. Input: "the local languages are studied by some linguists"
> Output:
> ...

The prompt with Italian input sentence, and requesting an Italian functional version is shown below.

set-up	ident	struct	pron
En-It 1-shot	0	0.63	0.24
En-It 5-shot	0	0.66	0.48
It-It 1-shot	0.03	0.76	0.79
It-It 5-shot	0.08	0.79	0.83

Table 2
Testing English as a "pivot language" for the LLaMa generative model. Transforming an English input sentence into the Italian functional form (En-It) and the Italian sentence into its functional form (It-It).

> Replace noun phrases with pronouns and prepositional phrases with adverbs. Preserve the exact syntactic structure, word order, and verb forms.
>
> Examples:
>
> Input: "i suoi giocattoli erano intagliati dai suoi genitori nella baita" → Output: "questi erano intagliati da loro là"
>
> Now convert these:
>
> 1. Input: "quella canzone era canticchiata dai miei amici da qualche settimana"
> Output:
>
> 2. Input: "le lingue del luogo sono studiate da alcuni linguisti"
> Output:
> ...

3.2.2. Evaluation

To evaluate the outputs, we use three complementary measures: (i) *perfect match* (ident) the percentage of instances for which the system generation matches the gold standard (ii) *structure match* (struct), for each output we compute an F1 score that quantifies how well the system has predicted the structure [5] and (iii) *pronoun/adverb ratio* (pron), where we compute the ratio of pronouns and adverbs in the system output and the pronouns and adverbs in the gold standard. All these measures are rough approximations, and overestimate the performance, but in a consistent manner. Table 2 shows these measures for the four experimental set-ups.

Similarly to the experiments on the monolingual and multilingual encoder models, the experiments on the generative LLM has shown that forcing multiple languages to share the parameter space leads to the loss of syntactic, semantic and lexical language-specific information. The set-up does not lead to the encoding of shared abstract grammatical representations [1, 3, 19, 4, 5]. Whether English is the internal language of generative LLMs from the LlaMa family or not [7, 8], the structure of English sentences does not seem to be better encoded than for Italian. Furthermore, the match between the language of the input and the output seems to be of importance.

[5] We obtain dependency relations for the system output and the gold standard using spaCy (https://spacy.io/v.3.8.7), and computed the F1 based on the true positive count (how many relations overlap), false positive (how many additional relations the system answer has relative to the gold standard) and false negatives (how many dependencies the gold standard has that do not appear in the system output).

4. Discussion

We aimed to explore the impact of encoding together multiple language, with English dominating the training data, for encoder and decoder language models.

The comparison of detecting syntactic-semantic structure using a multilingual and a monolingual encoder model has shown that the monolingual Italian model encodes both structural and linguistic abstraction information in a cleaner and more accessible way compared to a multilingual model, contrary to previous hypotheses about multilingual training leading to the encoding of more abstract linguistic structures. We have shown this effect through an exploration of individual sentences, as well as when the sentence structure was required to solve a more complex linguistic puzzle. Adding the lexical abstraction level to the structure exploration allows us to reach the shared structures of lexicalized and functional sentence variations.

Using a decoder transformer model, we have explored sentence structure encoding through the generative lens: how well does a system recognize and preserve the syntactic and semantic structure of an input sentence. Because it has been shown that English functions as a latent language, it would be expected that the structure of an English sentence is more readily detected and preserved. We found that that is not the case, and mapping a lexicalized Italian input sentence into its functional form leads to better results, both in terms of preserving the structure, and in the generation of pronominal and adverbial replacements for noun and prepositional phrases.

5. Related work

Multilingual models project many languages in the same parameter space. This brings some clear advantages: the model can be moved easily between different language applications, and it allows for low-resource languages to be bootstrapped by their connections to other languages. It has been surmised that forcing multiple languages to share the same parameter space will lead to the emergence of linguistic universals. It has been shown that that LLMs generalize across languages through implicitly learned vector alignment, which is less robust for generative models [20]. Some work using cross-lingual

structural priming finds evidence that grammatical representations are abstract and shared in multilingual language models [5] . Further exploration has found, however, that this effect depends on the similarity between the included languages [21]. It has also been shown that models encode grammatical information, such as chunks and structure, in a language-specific manner [6]. Overall, it is difficult to draw a conclusion on the performance of multilingual models, because it can be overestimated due to skewed language selection [22].

There are also downsides to building a multilingual model, as language particularities may be lost in the shared space, particularly when there is a dominant language. This may lead to language confusion in generation [23], and a decrease in the faithfulness of the multilingual models compared to monolingual ones, assessed in terms of feature attribution [24]. An asymmetrical effect of recall in monolingual and multilingual models depending on the syntactic role (subject vs. object) has also been found [25]. Finally, the language of the prompt affects a multilingual model's performance on binary questions about sentence grammaticality [26].

6. Conclusions

The current work aimed to explore the costs or advantages of multilingual and monolingual models, in a linguistic problem that involves a form of abstraction in language models. In particular, we focused on the issue of lexical abstraction through functional words – pronouns and adverbs standing in for noun and prepositional phrases. Lexicalized and functional versions of the same sentence share syntactic structure and semantic roles, information which should be encoded by language models. We tested whether this information is identifiable and whether lexicalized and functional parallel sentences encode this information in a similar manner. We explored multilingual models, testing the assumption that forcing many languages to share the same parameter space leads to a more abstract encoding of information. We found that this assumption does not hold in either encoder or decoder models.

Acknowledgments

We gratefully acknowledge the support of this work by the Swiss National Science Foundation, through grant SNF Advanced grant TMAG-1_209426 to PM.

References

[1] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[2] S. Wu, M. Dredze, Are all languages created equal in multilingual BERT?, in: S. Gella, J. Welbl, M. Rei, F. Petroni, P. Lewis, E. Strubell, M. Seo, H. Hajishirzi (Eds.), Proceedings of the 5th Workshop on Representation Learning for NLP, Association for Computational Linguistics, Online, 2020, pp. 120–130. URL: https://aclanthology.org/2020.repl4nlp-1.16/. doi:10.18653/v1/2020.repl4nlp-1.16.

[3] A. Conneau, S. Wu, H. Li, L. Zettlemoyer, V. Stoyanov, Emerging cross-lingual structure in pretrained language models, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6022–6034. URL: https://aclanthology.org/2020.acl-main.536/. doi:10.18653/v1/2020.acl-main.536.

[4] A. Sinclair, J. Jumelet, W. Zuidema, R. Fernández, Structural persistence in language models: Priming as a window into abstract language representations, Transactions of the Association for Computational Linguistics 10 (2022) 1031–1050. URL: https://aclanthology.org/2022.tacl-1.60/. doi:10.1162/tacl_a_00504.

[5] J. Michaelov, C. Arnett, T. Chang, B. Bergen, Structural priming demonstrates abstract grammatical representations in multilingual language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 3703–3720. URL: https://aclanthology.org/2023.emnlp-main.227/. doi:10.18653/v1/2023.emnlp-main.227.

[6] V. Nastase, G. Samo, C. Jiang, P. Merlo, Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 631–643. URL: https://aclanthology.org/2024.clicit-1.71/.

[7] C. Wendler, V. Veselovsky, G. Monea, R. West, Do llamas work in English? on the latent language of multilingual transformers, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15366–15394. URL: https://aclanthology.org/2024.acl-long.820. doi:10.18653/v1/2024.acl-long.820.

[8] I. Papadimitriou, K. Lopez, D. Jurafsky, Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models, in: A. Vlachos, I. Augenstein (Eds.), Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1194–1200. URL: https://aclanthology.org/2023.findings-eacl.89/. doi:10.18653/v1/2023.findings-eacl.89.

[9] G. Samo, A structured synthetic dataset of English and Italian verb alternations for testing lexical abstraction via functional lexicon in LLMs, 2025. URL: https://ling.auf.net/lingbuzz/009085. arXiv:lingbuzz/009085, preprint available at lingbuzz/009085.

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

[11] P. Merlo, S. Stevenson, Automatic verb classification based on statistical distributions of argument structure, Computational Linguistics 27 (2001) 373–408. URL: https://aclanthology.org/J01-3003/. doi:10.1162/089120101317066122.

[12] G. Samo, A structured synthetic dataset of English and Italian verb alternations for testing lexical abstraction via functional lexicon in LLMs, 2025. URL: https://ling.auf.net/lingbuzz/009085, preprint available at lingbuzz/009085.

[13] 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.

[14] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[15] 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.

[16] V. Nastase, P. Merlo, Are there identifiable structural parts in the sentence embedding whole?, in: Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, H. Chen (Eds.), Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Miami, Florida, US, 2024, pp. 23–42. URL: https://aclanthology.org/2024.blackboxnlp-1.3/. doi:10.18653/v1/2024.blackboxnlp-1.3.

[17] 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.

[18] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, C. Olah, Toy models of superposition, 2022. URL: https://arxiv.org/abs/2209.10652. arXiv:2209.10652.

[19] A. Jones, W. Y. Wang, K. Mahowald, A massively multilingual analysis of cross-linguality in shared embedding space, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5833–5847. URL: https://aclanthology.org/2021.emnlp-main.471/. doi:10.18653/v1/2021.emnlp-main.471.

[20] Q. Peng, A. Søgaard, Concept space alignment in multilingual LLMs, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 5511–5526. URL: https://aclanthology.org/2024.emnlp-main.315/. doi:10.18653/v1/2024.emnlp-main.315.

[21] C. Arnett, T. A. Chang, J. A. Michaelov, B. Bergen, On the acquisition of shared grammatical representations in bilingual language models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 20707–20726. URL: https://aclanthology.org/2025.acl-long.1010/. doi:10.18653/v1/2025.acl-long.1010.

[22] E. Ploeger, W. Poelman, M. de Lhoneux, J. Bjerva, What is "typological diversity" in NLP?, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 5681–5700. URL: https://aclanthology.org/2024.emnlp-main.326/. doi:10.18653/v1/2024.emnlp-main.326.

[23] K. Marchisio, W.-Y. Ko, A. Berard, T. Dehaze, S. Ruder, Understanding and mitigating language confusion in LLMs, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 6653–6677. URL: https://aclanthology.org/2024.emnlp-main.380/. doi:10.18653/v1/2024.emnlp-main.380.

[24] Z. Zhao, N. Aletras, Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3226–3244. URL: https://aclanthology.org/2024.naacl-long.178/. doi:10.18653/v1/2024.naacl-long.178.

[25] C. Fierro, N. Foroutan, D. Elliott, A. Søgaard, How do multilingual language models remember facts?, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 16052–16106. URL: https://aclanthology.org/2025.findings-acl.827/. doi:10.18653/v1/2025.findings-acl.827.

[26] S. Behzad, A. Zeldes, N. Schneider, To ask LLMs about English grammaticality, prompt them in a different language, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15622–15634. URL: https://aclanthology.org/2024.findings-emnlp.916/. doi:10.18653/v1/2024.findings-emnlp.916.

A. Blackbird Language Matrices data

Verb split between train and test for the COS and OD subsets. For the sentence representation analysis, the data respects the same split.

A.1. Data split

Table 3 below shows the train:test split of the 30 verbs for each of the change-of-state and object-drop verbs. 100 instances for each verb will be included completely either in the training or the test subsets.

Class	**Verb**	
	Train	**Test**
COS	addolcire, affilare, allargare, annerire, aprire, armonizzare, caramellare, chiudere, corrodere, cuocere, espandere, friggere, indurire, ingrandire, intensificare, migliorare, piegare, propagare, purificare, rimpicciolire, riscaldare, rompere, sbiancare, sciogliere, scongelare, stropicciare, svuotare	illuminare, scheggiare, strappare
OD	allattare, arare, bere, cantare, canticchiare, cucinare, cucire, dipingere, disegnare, giocare, impastare, insegnare, lavare, leggere, lucidare, mangiare, mungere, pescare, pulire, rammendare, recitare, saldare, scolpire, seminare, spazzare, studiare, tessere	intagliare, scrivere, stirare

Table 3
BLM data: train/test verbs grouped by class

A.2. BLM task instances for change-of-state verbs

Table 4 show a lexicalized and functional instance from the change-of-state verbs.

COS		
Context		
	Functional	*Lexical*
1	Loro friggevano quelle lì per noi	Le contadine friggevano delle uova per la serata
2	Loro friggevano quelle lì da poco	Le contadine friggevano delle uova da pochi minuti
3	Quelle lì erano fritte da loro per noi	Le uova erano fritte dalle contadine per la serata
4	Quelle lì erano fritte da loro da poco	Le uova erano fritte dalle contadine da pochi minuti
5	Quelle lì erano fritte per noi	Le uova erano fritte per la serata
6	Quelle lì erano fritte da poco	Le uova erano fritte da pochi minuti
7	Quelle lì friggevano per noi	Le uova friggevano per la serata
8	?	?
Answer Set		
1	**Quelle lì friggevano da poco**	**Le uova friggevano da pochi minuti**
2	Loro friggevano da poco	Le contadine friggevano da pochi minuti
3	Quelle lì erano fritte da loro	Le uova erano fritte dalle contadine
4	Loro erano fritte da quelle lì	Le contadine erano fritte dalle uova
5	Quelle lì friggevano loro	Le uova friggevano le contadine
6	Loro friggevano quelle lì	Le contadine friggevano le uova
7	Quelle lì friggevano da loro	Le uova friggevano dalle contadine
8	Loro friggevano da quelle lì	Le contadine friggevano dalle uova

Table 4
Example for ItCOSFun and ItCOSLex

A.3. BLM task instances for object-drop verbs

Table 5 show a lexicalized and functional instance from the object-drop verbs.

OD		
Context		
	Functional	*Lexical*
1	Lei recitava questa per loro	L'artista recita una poesia in fiorentino antico
2	Lei recitava questa da qui	L'artista recita una poesia da qualche giorno
3	Questa era recitata da lei per loro	La poesia è recitata dall'artista in fiorentino antico
4	Questa era recitata da lei da qui	La poesia è recitata dall'artista da qualche giorno
5	Questa era recitata per loro	La poesia è recitata in fiorentino antico
6	Questa era recitata da qui	La poesia è recitata da qualche giorno
7	Lei recitava per loro	L'artista recita in fiorentino antico
8	?	?
Answer Set		
1	Questa recitava da qui	La poesia recita da qualche giorno
2	**Lei recitava da qui**	**L'artista recita da qualche giorno**
3	Questa era recitata da lei	La poesia è recitata dall'artista
4	Lei era recitata da questa	L'artista è recitata dalla poesia
5	Questa recitava lei	La poesia recita l'artista
6	Lei recitava questa	L'artista recita la poesia
7	Questa recitava da lei	La poesia recita dall'artista
8	Lei recitava da questa	L'artista recita dalla poesia

Table 5
Example for ItODFun and ItODLex