

Comparing Natural and Synthetic Structured Data: A Study of the Passive Verb Alternation in French and Italian

Giuseppe Samo¹, Paola Merlo^{1,2}

¹IDIAP Research Institute, ²University of Geneva
{giuseppe.samo, paola.merlo}@idiap.ch

Abstract

This study compares the impact of natural and synthetic data on training and evaluating large language models (LLMs), using the case of passive verb alternation in French and Italian. We use Blackbird Language Matrices (BLMs), structured datasets designed to probe linguistic knowledge of underlying patterns across sentence sets. We compare structured templates instantiated with natural sentences extracted from Universal Dependencies to structured templates of synthetic sentences. Experiments show that while models achieve ceiling performance when trained and tested on synthetic datasets, they do not reliably generalize to natural sentences. In contrast, models trained on natural data exhibit robust performance across both natural and synthetic test suites, demonstrating their superior ability to capture abstract linguistic patterns. These results corroborate the value of natural data and of structured set ups in linguistic evaluation for probing LLMs' syntactic and semantic knowledge.

Keywords: Verb alternation, Universal Dependencies, Language Models, French, Italian

1. Introduction

Whether to use natural or synthetic data for evaluating linguistic knowledge has long been a central methodological debate in theoretical linguistics (Chomsky, 1965; Gibson and Fedorenko, 2013) and computational linguistics (cf. *test suites* vs. *test corpora*; Lehmann et al. 1996).

Natural data are easy to define: they are naturally occurring instances – ‘wild data’ (Bresnan, 2016) – extracted from corpora and can be annotated with various types of linguistic information. Synthetic data, by contrast, are artificially generated examples designed to target specific linguistic phenomena under controlled conditions. In this sense, synthetic datasets are conceptually akin to experimental materials in psycholinguistics, where carefully controlled stimuli are constructed to isolate particular grammatical factors (Schütze et al., 2013; Sprouse and Almeida, 2017; Futrell et al., 2019).

To ground this contrast, consider a simple verb alternation, where a verb can appear in a transitive or intransitive form (see also, unspecified object alternation, Levin 1993), such as the Italian verb *cantare* ‘sing’. We can colour-code the core elements of the argument structure by marking the constituent with the **Agent** role in purple, the **verb** in teal, and the **Theme** in blue. In a synthetic setting, one can construct minimal, grammar-like examples that clearly instantiate the alternation and control the relevant variables. Consider, for example, the contrast between *La ragazza canta la canzone* (the girl is singing the song) and *La ragazza canta* (the girl is singing). The same structural configurations and their comparison, however, can also

be retrieved from naturally occurring corpora by querying dependency relations such as `nsubj` and `obj` in annotated treebanks of Universal Dependencies (henceforth UD, Nivre 2015; De Marneffe et al. 2021). For example, consider the naturally occurring sentences below, which instantiate the same relevant constructions: ‘*Mina ha anche cantato molte canzoni di Mogol*’ (Mina [a singer] has sung many songs of [written by] Mogol [an author] (ISDT, 2Parole_6-47)) and ‘*inoltre sino all’ultimo non si sapeva se il soprano ungherese Andrea Rost, ammalatasi nei giorni scorsi, avrebbe potuto cantare nel ruolo della protagonista Violetta, come alla fine è stato*’ (moreover, until the very end it was not known whether the Hungarian soprano Andrea Rost, who had fallen ill in the previous days, would be able to sing the role of the protagonist Violetta, as in the end she did (ISDT, isst_tanl-1773)).

The notion of synthetic data has broadened considerably in recent years. It includes devised novel examples for experimental purposes, and it also refers to automatically generated outputs from language models and other conversational AI systems (Schepens et al., 2025; Civico, 2025). Such data may be synthetic in multiple senses: they are generated automatically rather than naturally occurring, and they can be tailored for experimental purposes. For example, Figure 1 shows an interaction in English with DeepSeek V.3. The interaction creates a synthetic instantiation for the alternation of the Italian verb *cantare* by prompting the generation of linguistic examples in English.¹ Recent work in NLP has increasingly explored the use of syn-

¹For comparison, ChatGPT (GPT-5-mini, OpenAI) generated the following: “Intransitive use: *Ogni mattina Maria canta sotto la doccia*. (Here, *cantare* has no direct

Human:

Can you make two sentences that show the fact that the verb *cantare* in Italian can be both transitive and intransitive?

DeepSeek:

Certainly! Here are two sentences that demonstrate *cantare* used transitively (with a direct object) and intransitively (without a direct object):

Intransitive use: *Maria canta sotto la doccia ogni mattina. (Here, cantare is used without a direct object—Maria is simply singing.)*

Transitive use: *Maria canta una bellissima canzone d'amore. (Here, cantare takes a direct object—una bellissima canzone d'amore.)*

Figure 1: Example of a synthetic piece of data for transitive and intransitive forms of the Italian verb *cantare* ‘sing’ generated by DeepSeek (DeepSeek-AI, accessed Feb 13, 2026).

thetic data to complement or even replace natural data in training and evaluation pipelines (Goyal and Mahmoud, 2024; Josifoski et al., 2023; Chen et al., 2024). However, as noted by Jumelet et al. (2025), although large-scale annotated resources such as UD have become somewhat marginalized in the era of large language models (LLMs), the extensive linguistic knowledge encoded in these datasets continues to provide an essential foundation for rigorous and linguistically informed evaluation of current language models (Opitz et al., 2025).

This paper contributes to this debate by systematically comparing natural and synthetic data within a highly-structured diagnostic test suite targeting specific linguistic phenomena. We examine the extent to which each type of data helps language models identify and abstract the core linguistic elements underlying a complex phenomenon such as a verb alternation—one that cannot be adequately captured by considering sentences in isolation.

We construct structured datasets for the Blackbird Language Matrices (BLM) task (Merlo, 2023b,a; An et al., 2023; Samo et al., 2023; Nastase et al., 2024a,b; Jiang et al., 2024; Samo and Merlo, 2026a). The BLM task, inspired by Raven’s Progressive Matrices (Raven, 1938), presents models with sentence-based puzzles designed to capture underlying linguistic rules. Specifically, each

object—Maria is simply singing.); Transitive use: *Maria canta una canzone napoletana.* (Here, *cantare* takes a direct object—una canzone napoletana.). Accessed Feb 13, 2026.

task instance consists of a context set of sentences that implicitly encodes the information required to complete a given linguistic pattern, and an answer set of minimally contrasting candidates, only one of which correctly replaces the missing element in the pattern defined by the context. This general template is then instantiated using carefully curated datasets. We instantiate a BLM template in two ways: with natural sentences extracted from UD and with synthetic sentences generated with the help of a ConversationalAI, each in two Romance languages, French and Italian.

The core linguistic phenomenon under investigation is the passive alternation (Levin, 1993, 85) because of its general applicability, across lexical items and languages. Unlike other verb alternations, the passive is not strongly tied to a specific lexical class, but most transitive verbs can undergo passivization (e.g. in Italian with the passive voice – colour coded in orange – as in *La canzone è cantata dalla ragazza* ‘the song is sung by the girl’), making it a structurally defined alternation, widely attested cross-linguistically (Siewierska, 2013).

Our contribution is two-fold: (i) we introduce multilingual, structured datasets that systematically compare natural and synthetic data with the BLM task and (ii) we evaluate how these data types may shape model representations and their capacity to generalize. This study sheds light on the interaction between training data for structured test suites and linguistic generalization, supporting Jumelet et al. (2025)’s intuition that natural data constitute a robust basis for evaluating language models.

2. Modelling the Passive Alternation with BLM Paradigms

Verb alternations have long been a central object of study in theoretical and computational linguistics, as they provide a window into how syntax, semantics, the lexicon, and the functional lexicon interact (Levin, 1993; Rappaport Hovav and Levin, 2024). However, many alternations that are commonly investigated are lexically specific, only restricted verb classes participate, and cross-linguistic correspondences are often unstable (Fillmore et al., 2003). This lexical specificity can favour superficial patterns in pretrained language models, which may capture surface-level lexical associations rather than abstract syntactic generalizations. The passive alternation, instead, is less restricted in the lexical nature of the verb it can apply to.

Passive constructions preserve the core argument structure while systematically altering the mapping between syntactic positions and discourse roles (Hopper and Thompson, 1980). The core argument structure of the verb is preserved, but its syntactic realization changes. The verb’s **Theme** is

promoted to subject position and, in languages with overt agreement, controls subject–verb agreement, while the **Agent** can be omitted from the surface form, yielding structures that are underlyingly diadic (like the transitive sentence in the introduction) but appear monadic on the surface (like the intransitive sentence in the introduction).

Across languages, passives are typically signaled by changes in word order and functional morphosyntactic markers. In Italian and French, passive constructions are mainly marked by the presence of an auxiliary verb. Italian typically uses *essere* ‘be’ or *venire* ‘come’, while French primarily uses *être* ‘be’ and, in certain constructions, also *faire* ‘do’. However, these auxiliaries are not specific to passive structures. For example, in both languages *essere* and *être* ‘be’ also introduce copular and predicative constructions; in Italian, *essere* is additionally used to form compound tenses with certain classes of verbs such as unaccusatives (Burzio, 1986). In French, *faire* ‘do’ can also be used in causative constructions. Finally, *venire* ‘come’ – as well the other auxiliaries – function as lexical verbs that can occur in isolation.

The passive form is particularly relevant, as it constitutes a highly pervasive phenomenon in Romance, especially in specific text genres (Brunato et al., 2022), while at the same time displaying considerable structural complexity (Volpato, 2010) and giving rise to effects at the pragmatic level (Reinhart, 1981, 64). As a result, passivization is less fine-grained, and also differently generalizable, than other lexical alternations.

As a key methodological choice in our study of alternation phenomena, we move beyond isolated sentences as the primary unit of analysis. Single sentences are often insufficient, since alternations are inherently relational; they concern systematic correspondences between forms rather than properties of individual strings (unlike, for instance, agreement or long-distance dependencies). Minimal pairs are also a well-established tool in linguistic analysis, but they rely on tightly controlled contrasts and inevitably remain tied to specific lexical items.

For this reason, we adopt a paradigm-based perspective, where the primary object of analysis is a structured set of systematically related constructions. We chose to evaluate the knowledge of passive with the BLM task, which has been shown to be one of the most challenging evaluations for testing Italian knowledge in LLMs (Nissim et al., 2025).

The BLM template The template adopted in this study is shown in Figure 2. BLM templates are typically constructed around three core elements: external rules E , internal rules I and relational operators R (Merlo, 2023b). The external rule E of

Template			
Structure	R_{Arg}	R_{ST}	Example
CONTEXT SET			
1 Ag Vact Th ?	2	Q	<i>Does the customer pay the bill?</i>
2 Ag Vact Th	2	D	<i>The student gets the prize.</i>
3 Ag Vact ?	1	Q	<i>Does the teacher explain?</i>
4 Ag Vact	1	D	<i>The car moves.</i>
5 Th Vpass Ag ?	2	Q	<i>Why is the case studied by the lawyer?</i>
6 Th Vpass Ag	2	D	<i>The key is found by the boy.</i>
7 Th Vpass ?	1	Q	<i>When was the screen touched?</i>
8 ???			
ANSWER SET			
1 Th Vpass	1	D	<i>The plants were watered</i>
2 Th Vpass Ag	2	D	<i>The news is reported by the speaker.</i>
3 Ag Vact	1	D	<i>The writer publishes.</i>
4 Ag Vact Th	2	D	<i>The store ships the order.</i>
5 Question	1 or 2	Q	<i>How was the data analyzed?</i>

Figure 2: BLM template structure instantiated with a synthetic example in English, generated with DeepSeek V.3 (section 3.2). **Ag** = Agent, **Th** = Theme, **Vact** = verb in active voice, **Vpass** = verb in passive voice, **red** elements mark interrogative markers. Number of arguments (1, 2) and sentence type (Q = question, D = declarative).

the phenomenon that distinguishes passive from active voice concerns verbal morphology. The internal rule I relates to the syntax–semantic mapping of the verb’s arguments. In passive constructions, the **Theme** is promoted to syntactic subject position and agrees with the verb. The R components specifies how the linguistic objects of the BLM are connected within the template. In the proposed template, R corresponds to sentence type (R_{ST}) and number of arguments superficially realized in the sentence (R_{Arg}). These two linguistic phenomena are external to the one under investigation and do not influence the learning of E and I rules. We explore the contrast between interrogative and declarative sentences, but also between bi-argumental and mono-argumental structures.

The inclusion of both declarative and interrogative clauses is partly motivated by the structure of the available treebanks and the distributions of sentences across treebanks. For example, the French QuestionBank (FQB, Seddah and Candito 2016) exclusively contains interrogative sentences, and allows the retrieval of a large number of instances for task training and testing purposes. Declaratives are also present in large scale in the UD treebanks of French and Italian (Samo and Merlo, 2021). From a linguistic perspective, sentence type constitutes an independent dimension with respect to passive voice alternation. Interrogative marking may affect word order, but does not interfere with the morphological marking of passive voice (E), nor with the argument-structure reconfiguration involved in the promotion of the **Theme** (I).

We therefore develop a template that manipulates the organization of the three core elements E ,

I, and *R*. Elements of *E* include the verb inflected for the active voice (*Vact*) and for the passive voice (*Vpass*). The *I* component concerns the syntactic functions of the core arguments, *Agents* and *Themes*. The *R* component specifies whether both *Agents* and *Themes* are present in the same string (or only one of these arguments appears, R_{Arg}) and whether the sentence is interrogative or declarative. Interrogative sentences are marked typographically by the interrogative symbol *?*, and also by the presence of *wh*-elements (R_{ST}).

Context Set Formally, the context set instantiates a $2 \times 2 \times 2$ design crossing voice (active/passive), number of argument realizations (one or two overt arguments), and sentence type (interrogative vs. declarative). The eighth sentence is left uninstantiated. The organization of the context reflects two independent alternations built on the different *R*s (R_{Arg} and R_{ST}).

Each pair of consecutive sentences forms a minimal interrogative–declarative paradigm: in all four pairs, the first member is interrogative and the second is declarative. For every four sentences, the first two show two overt arguments, while the last two show only one argument. The resulting structure creates a double alternation pattern (and a triple one with active-passive): sentence type alternates within each pair, while voice and argument realization alternate across pairs.

The first pair (1–2) consists of active clauses with two overt arguments. The second pair (3–4) maintains the active voice, but in mono-argumental structures. The third pair (5–6) shifts to the passive voice while restoring the bi-argumental configuration, with the *Theme* promoted to subject position and the *Agent* realized as an overt agentive complement. Finally, the fourth pair is designed to instantiate mono-argumental passive clauses; however, only the interrogative configuration (7) is provided in the context. The corresponding declarative passive without an overt agent (8) is omitted and must be inferred. The missing final sentence can therefore be recovered only by integrating all the patterns simultaneously.

Answer Set The correct answer is the declarative passive sentence without an agentive prepositional phrase. The answer set is designed to cover all combinations of argument number (one or two arguments) and voice (active vs. passive), and an additional interrogative option. Therefore, the errors are of three types: (i) errors of *VOICE* when the verb is inflected in the active voice, violating *E* and implicitly *I*; (ii) errors of *NUMBER OF ARGUMENTS* when there are two arguments; (iii) errors of *SENTENCE TYPE* when the sentence is not a declarative, both violations of *R*. In our case, candidate

answer 2 (*Th Vpass Ag*) is an error of *NUMBER OF ARGUMENTS* because it also shows an additional argument, the agentive PP; sentence 3 (*Ag Vact*) is an error of *VOICE* because it is a mono-argumental declarative, but not a passive structure; sentence 4 (*Ag Vact Th*) is an error of both *VOICE* and *NUMBER OF ARGUMENTS*; sentence 5 (*question*) is mainly an error of *SENTENCE TYPE*.

3. Natural and Synthetic Data for BLM

The template illustrated in Figure 2 is instantiated with natural and synthetic data. Examples are given in French and Italian respectively in Figure 3 and Figure 4.²

3.1. Natural data

All the sentences are extracted from UD treebanks. We retrieved natural occurring sentences with the system GREWMATCH (MATCH.GREW.FR) (Guillaume, 2021). For French, we extracted questions from the French QuestionBank v.2.17 (FQB, 23,345 tokens; Seddah and Candito 2016) containing only questions from news and nonfiction sources. We extracted declarative sentences from the French GSD data v.2.17 (389,364 tokens; Guillaume et al. 2019), which contains data from blogs, news, reviews and encyclopedic entries. For Italian, we retrieved sentences from ISDT v.2.17 (278,424 tokens; Bosco et al. 2014) containing data from legal texts, news and Wikipedia and VIT v.2.17 (259,625 tokens; Alfieri and Tamburini 2016), which consists of news and nonfiction texts.

The sentence types of the BLM template were directly instantiated using sentences that matched the corresponding structures in the treebanks. Each context and answer sentence was retrieved using simple queries. The *Agent* in the sentences was identified as the target of the *nsubj* dependency associated with a variable *V* in active sentences, and as the dependent of a *obl:agent* dependency in passive sentences. The *Theme* was identified as the dependent of *obj* in active sentences and of the *nsubj:pass* dependency in passive sentences. To retrieve questions, we added the query for an interrogative marker corresponding to a variable *Q* whose *FORM* was annotated as *"?"*. We restricted our search to sentences – intended here as the output of an instance of the query – with a single verb by excluding output sentences containing additional verbs, using the constraint *Y* (a variable different than *V*) whose *UPOS* was *"VERB"*.

²Data is available at <https://www.idiap.ch/en/scientific-research/data/blm-passF> and <https://www.idiap.ch/en/scientific-research/data/blm-passI>.

Natural	CONTEXT SET	Synthetic
1	<i>Que signifie l'acronyme NASA ?</i> 'what does the acronym NASA mean?' (FQB_TREC-fr-584)	<i>Le garçon jette-t-il la pierre?</i> 'Does the boy throw the stone?'
2	<i>Leur album No Dice obtient un excellent accueil.</i> 'Their album No Dice obtained an excellent reception' (fr-ud-train_06225)	<i>L'équipe félicite le gagnant.</i> 'The team congratulates the winner'
3	<i>Combien de temps dure le voyage entre Tokyo et Niigata ?</i> 'How long does the journey from Tokyo to Niigata take?' (FQB_TREC-fr-111)	<i>L'équipe coûte-t-elle?</i> 'Does the team cost?'
4	<i>Bryon Anthony McCane naît en 1976 d'une mère italienne et d'un père afro-américain.</i> 'Bryon Anthony McCane was born in 1976 to an Italian mother and an African-American father.' (fr-ud-train_00556)	<i>Le chanteur chante.</i> 'The singer sings'
5	<i>Les moteurs rotatifs étaient produits par qui ?</i> 'By whom were rotating engines invented?' (FQB_TREC-fr-886)	<i>Comment la scène est-elle décrite par l'écrivain ?</i> 'How is the scene described by the writer?'
6	<i>Uber a été fondé par Garrett Camp, Travis Kalanick et Oscar Salazar en 2009.</i> 'Uber was founded by Garrett Camp, Travis Kalanick and Oscar Salazar in 2009' (fr-ud-train_02964)	<i>Un livre est écrit par l'auteur.</i> 'A book is written by the author'
7	<i>Dans quelle province française le cognac est-il produit ?</i> 'In which French province is cognac produced?' (FQB_TREC-fr-1134)	<i>Quand la musique a-t-elle été composée?</i> 'When was the music composed?'
8	???	???
ANSWER SET		
1	Le château est ensuite vendu plusieurs fois 'The castle is then sold several times' (fr-ud-train_00010)	Les données ont été analysées. 'The data were analyzed'
2	<i>Les travaux furent dirigés par le Florentin Girolamo della Robbia et les Tourangeaux Pierre Gadier et Gatien François.</i> 'The works were directed by the Florentine Girolamo della Robbia and the Turingians Pierre Gadier and Gatien François' (fr-ud-train_07752)	<i>La langue est apprise par l'étudiant.</i> 'The language is learned by the student.'
3	En tout cas je n'y retourne pas et ma collègue non plus ! 'In any case, I'm not going back there and neither my colleague!' (fr-ud-train_00110)	<i>Le programmeur code.</i> 'The programmer codes.'
4	<i>En mars 2010, il signe son premier contrat professionnel avec Birmingham City.</i> 'In March 2010, he signed his first professional contract with Birmingham City' (fr-ud-train_01336)	<i>Le parent paie la facture.</i> 'The parent pays the bill'
5	<i>Quel âge a le Soleil ?</i> 'How old is the Sun?'(FQB_TREC-fr-334)	<i>Quand le colis a-t-il été reçu?</i> 'When was the package received?'

Figure 3: Natural and synthetic examples in French, glosses and ID number (natural data) for reference within the explored treebanks. We have coloured code the core elements of the BLM-template. Correct answer in bold.

Example Natural	CONTEXT SET	Example Synthetic
1	<i>Quando Innsbruck ospitò le Olimpiadi Invernali?</i> 'When did Innsbruck host the Winter Olympic Games?' (ISDT, quest-991)	<i>Lo scrittore finisce il romanzo?</i> 'Does the writer finish the novel?'
2	<i>I carabinieri gli hanno recapitato il decreto di revoca degli arresti domiciliari.</i> 'The Carabinieri delivered to him the order revoking his house arrest.' (ISDT, isst_tanl-280)	<i>Il giocatore segna un gol.</i> 'The player scores a goal.'
3	<i>A che velocità viaggia la luce?</i> 'At what speed does light travel?' (ISDT, quest-656)	<i>Il bambino disegna?</i> 'Does the child draw?'
4	<i>E nel frattempo politici, ambasciatori e industriali discutevano di petrolio.</i> 'And in the meantime politicians, ambassadors, and industrialists were discussing oil.' (ISDT, isst_tanl-186)	<i>L'editore corregge.</i> 'The publisher corrects.'
5	<i>Chi è stato sconfitto da Andrei Medvedev nella finale del Torneo di Monte Carlo ?</i> 'Who was defeated by Andrei Medvedev in the Monte Carlo tournament final?' (ISDT, quest-553)	<i>Quando sono annaffiate le piante dal giardiniere?</i> 'When are the plants watered by the gardener?'
6	<i>Le leggi sono promulgate dal Presidente della Repubblica entro un mese dall'approvazione</i> 'The laws are promulgated by the President of the Republic within one month of approval.' (ISDT, tut-1546)	<i>La squadra è tifata dal bambino.</i> 'The team is supported by the child.'
7	<i>Come può essere eliminato il rigetto del sistema immunitario?</i> 'How can immune system rejection be eliminated?' (quest-297)	<i>Quando è stato sollevato il peso?</i> 'When was the weight lifted?'
8	???	
ANSWER SET		
1	Dopo tre anni di discussioni la proposta non venne accolta. 'After three years of discussions, the proposal was not accepted.' (VIT-1304)	Lo strumento è stato suonato. 'The instrument was played.'
2	<i>La lettera di trasporto aereo viene emessa dal mittente in tre esemplari originali.</i> 'The air waybill is issued by the sender in three original copies.' (ISDT, splet-161)	<i>L'email è inviata dal capo.</i> 'The email is sent by the boss.'
3	<i>La squadra nazionale italiana di calcio giocherà la prima partita del Campionato mondiale il 12 giugno, ad Hannover, con il Ghana.</i> 'The Italian national football team will play its first World Cup match on June 12 in Hanover against Ghana.' (ISDT, 2Parole_1-214)	<i>Il cassiere conta i soldi.</i> 'The cashier counts the money.'
4	<i>I disegni risentono delle influenze arabe e a volte anche di quelle francesi</i> 'The drawings show Arab influences and sometimes French ones as well.' (ISDT, isst_tanl-3212)	<i>Lo chef cucina.</i> 'The chef cooks.'
5	<i>Chi possiede la Saab?</i> 'Who owns the Saab?' (ISDT, quest-422)	<i>L'autista parcheggia l'auto?</i> 'The driver parks the car?'

Figure 4: Natural and synthetic examples in Italian, glosses and ID number (natural data) for reference within the explored treebanks. We have coloured code the core elements of the BLM-template. Correct answer in bold.

The combination of these factors led to the retrieval of natural instances for each sentence of the context and the answer set.³

Samples of the naturally extracted data were manually checked to assess whether the queries returned the relevant structures.

3.2. Synthetic data

We created our dataset using the DeepSeek-V3 model (Guo et al., 2025), adopting a hybrid approach that combines LLM generation with explicit linguistic constraints. This ensured the generated sentences remained grammatically well-formed and semantically coherent within each alternation paradigm. Specifically, the prompt instructed the conversationalAI to generate a structured linguistic dataset consisting of "100 simple sentences with frequent subjects, verbs and objects" (5 interactions) in English and transform them into eight distinct syntactic variants corresponding to the context and answer set structures (interrogative, intransitive, and passive forms with and without agentive phrases), with specific formatting requirements for direct import into spreadsheet software. In later interactions, the model was required to translate the sentences into French and Italian. A native speaker of Italian and proficient in French and English, manually checked the translations. The intervention was minimal, as the sentences were generally well-formed.⁴

English was adopted as a pivot language (Wendler et al., 2024) to construct a single abstract dataset before its language-specific realization. This design choice ensures that the French and Italian versions do not stand in a derivational relationship to each other, but instead instantiate the same underlying synthetic templates. As a result, the synthetic status of the data is methodologically consistent across languages: neither dataset is more natural, or derived, than the other, since both originate from the same controlled generative procedure.

The sentences were specifically designed to isolate the factors under investigation: as shown in Figures 3 and 4, they included exclusively the elements relevant to the BLM (agents, themes, verbs in active and passive voice and interrogative markers).

³All queries are available in Table 1 in the Appendix.

⁴Out of 4,000 sentences (500 for each of the eight structures), only 43 were manually corrected for French (0.01) and 64 for Italian (0.02). The intervention primarily ensured that sentences remained unambiguously transitive or intransitive, removing verbs with reflexive-like elements when relevant for the template.

4. Experiments

The experiments systematically compare natural and synthetic data in a highly structured setup, the BLM test suite. The goal is to evaluate how much each type of data helps language models identify and abstract the core linguistic elements underlying the passive alternation. We explore the behavior of a simple probe through a series of experiments.

We adopt a feed-forward neural network (FFNN) architecture as described in Samo et al. (2023). For each sentence in the BLM, we compute an embedding by averaging its token representations obtained from pretrained models. The FFNN takes as input the concatenated embeddings representing the context, is trained using a max-margin loss objective, and predicts the answer whose embedding achieves the highest cosine similarity with the network's output. Following previous work (Nastase et al., 2025), we first test embeddings from monolingual ELECTRA models (French: *dbmdz/electra-base-french-europeana-cased-discriminator*; Italian: *dbmdz/electra-base-italian-xxl-cased-discriminator*) and then from a multilingual model (*google/electra-base-discriminator*). Monolingual and multilingual models may exhibit asymmetries in performance due to differences in their token representations (Samo and Merlo, 2026b).

We run both experiments when training and testing belong to the same data set type and also experiments that cross the two types. Each dataset, constituted of 2000 BLM instances, is split into separate training and testing sets (80%–20% split), ensuring no overlap of instances between them to control data leakage.

These settings are referred to as SYN_{SYN} (train and test on synthetic) and NAT_{NAT} (train and test on natural). SYN_{NAT} refers to training the model with synthetic data and testing its ability to generalize to natural instances, while NAT_{SYN} denotes training on BLM-templates instantiated with natural data and testing on its ability to abstract to synthetic instances.

4.1. Results: Monolingual

Figure 5 shows the results, measured in F1 scores, for the monolingual models. Among the monolingual models, the full synthetic suite achieves the highest performance (French 1.00; Italian 0.99), outperforming the natural suites (French 0.62; Italian 0.77).

When models are trained on synthetic data but tested on natural data, performance drops significantly, reaching near-chance levels (French 0.29; Italian 0.28). In contrast, training on natural data and testing on synthetic data yields performance comparable to that of the full natural suites, with

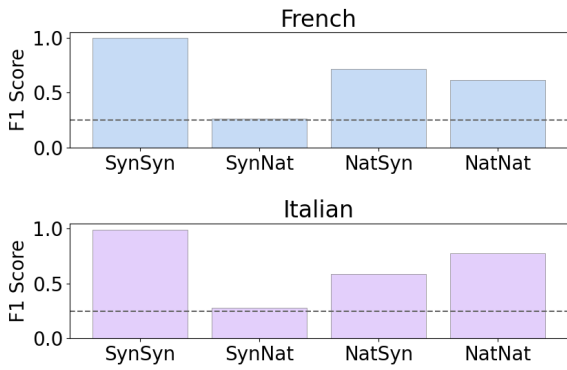


Figure 5: F1 scores across training and test suites in monolingual models. The grey dotted line indicates chance level.

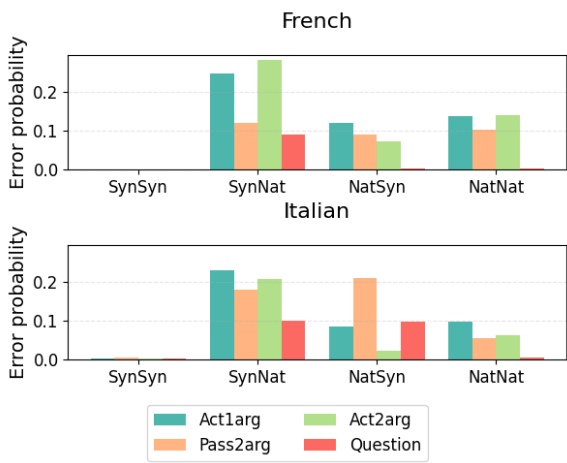


Figure 6: Error probabilities and types of errors (indicated by the selected erroneous answer) in monolingual models. Act = Active; Pass = Passive; arg = arguments.

French even showing improved performance with respect to natural test suites.

The error analysis is presented in Figure 6. Across all suites, errors related to SENTENCE TYPE (questions) are never the most prominent while the other errors are more evenly distributed. However, in the Italian NATSYN, the most prominent error type involves passive structures with two arguments, suggesting that voice was a learned feature.

4.2. Comparing Monolingual and Multilingual Embeddings

Figure 8 shows the results of the comparison between monolingual and multilingual models. The results follow the same pattern across languages and models. Multilingual models show competitive, and in some cases superior performance. In particular, the multilingual model reaches ceiling performance in the NATSYN suite for French (0.99),

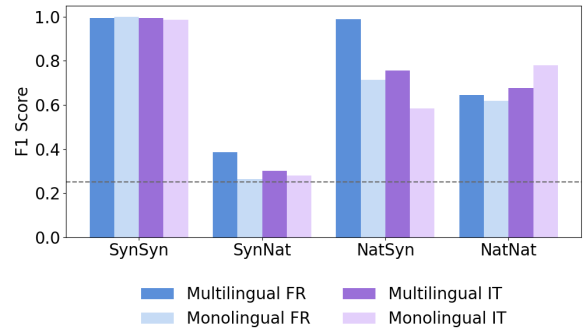


Figure 7: F1 scores across conditions, languages and models. The gray dotted line indicates chance level. IT = Italian, FR = French.

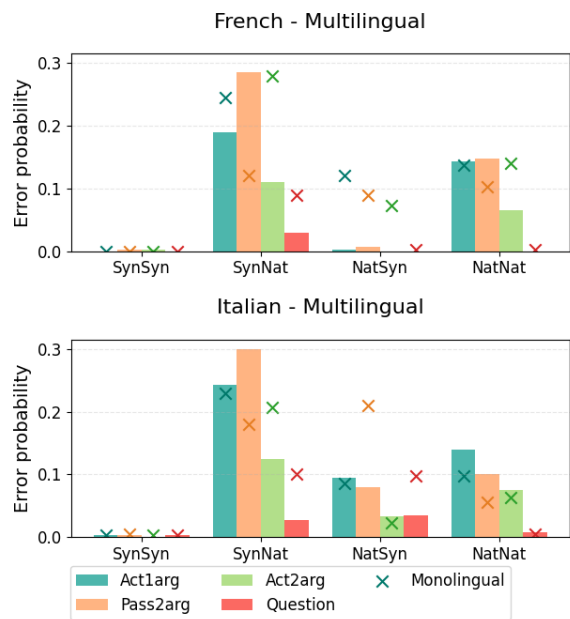


Figure 8: Error probabilities and types of errors (indicated by the selected erroneous answer) in multilingual models. Crosses repeat values of error bars in monolingual model for comparison (see Figure 6). Act=Active, Pass=Passive, arg=arguments.

indicating strong comprehension of the BLM and generalisation ability from natural to synthetic instances.

Consider the error analysis. In the NATNAT setting the most frequent incorrect responses in the multilingual models involve VOICE or NUMBER OF ARGUMENTS. This pattern indicates that active sentences with two arguments (i.e., violations of two rules) are not considered plausible answers. Errors related to SENTENCE TYPE are minimal in the multilingual models, even in the lowest-performing configuration. This suggests that the models consistently learn the sentence type encoded in the BLM template across languages and training con-

ditions.

However, multilingual and monolingual models differ in the type of error that is most prominent in the lowest-performing setting, SYN_{NAT}. In both languages, the most frequent errors produced by the multilingual models are passive structures, suggesting that these models successfully learn verb voice from synthetic training data. In contrast, monolingual models show a higher proportion of active constructions as their most prominent error.

5. Discussion

Although the full synthetic suites outperform the full natural suites, overall performance drops significantly when models are trained on synthetic data but tested on natural data (aggregated, *t*-test: $t(6) = 27.07$, $SE = 0.025$). Although models trained on natural data show lower absolute performance, their results are more stable across different evaluation settings.

Despite the fact that models trained and tested on synthetic data achieve near-ceiling performance, this competence does not reliably generalize the learnt structure to natural data. Training on synthetic data may bias the model toward artifacts introduced by the synthetic constructions themselves. As a consequence, models fail to generalize to natural test suites, which are structurally and contextually more complex—where complexity refers to the presence of additional constituents that are not directly relevant to the task. In contrast, models trained on natural data show more robust cross-condition generalization, performing competitively not only on natural test suites but also on synthetic ones.

Natural sentences—here extracted from UD treebanks—exhibit greater lexical variability (e.g. proper nouns or hapaxes) and additional linguistic objects surrounding the target syntactic configuration. Rather than hindering learning, however, this variability appears to promote abstraction of the core elements of the task, although not reaching ceiling performance in the testing suites.

The Universal Dependencies (UD) framework—as a particularly curated and cross-linguistically valid example of structured and annotated data in general—plays a central methodological role in this study. By extracting instances through dependency relations and morphological features, we were able to construct datasets that are both naturally occurring and structurally controlled. Crucially UD allows the same query to be executed consistently across different languages and text genres. This makes it possible to retrieve comparable structures in a uniform way, without redefining language-specific search criteria each time. In this sense, UD and structured annotations in general function as a

bridge between raw language data and the need to build theoretically motivated, structured diagnostics paradigms for natural data.

6. Related Work

Language models have demonstrated good performance on lexically restricted verb alternations and thematic roles, especially in English (Kann et al., 2019; Warstadt et al., 2019; Wilson et al., 2023; Samo et al., 2023; Proietti et al., 2022). Our work expands computational approaches to the passive alternation (Sasano et al., 2013; Leong and Linzen, 2023).

BLMs share objectives with datasets relying on synthetic or carefully controlled materials, designed to target specific grammatical phenomena under controlled conditions, often using minimal pairs or carefully crafted paradigms to probe the linguistic knowledge of large language models across languages (Warstadt et al., 2019, 2020; Xiang et al., 2021; Suijkerbuijk et al., 2025). Unlike datasets of minimal pairs, though, the BLM puzzle presents a complex structure and expands the minimal contrasts to many dimensions of variation.

While synthetic data allow broad and systematic coverage, they can introduce distributional biases, encouraging models to rely on superficial regularities rather than abstract rules (Nadăș et al., 2025; Griffiths et al., 2024). Our results partly resonate with the findings of Zhang and Pavlick (2025), who show that models trained on synthetic data may continue to rely on superficial heuristics that generalize poorly to targeted evaluations, even when overall benchmark performance improves.

Natural data drawn from annotated and structured resources provide rich linguistic information at the word (Batsuren et al., 2022) or sentence level (Nivre, 2015; De Marneffe et al., 2021), offering authentic variation and contextual richness. Such data allow researchers to systematically probe language models, designing tasks that implicitly target complex grammatical phenomena, including agreement (Jumelet et al., 2025) and argument structure alternations in morphologically-rich languages (Samo and Merlo, 2026b). By bridging naturalistic data and controlled evaluation, structured corpora provide a principled framework for investigating the inner workings of language models.

7. Conclusion

Our study shows that while synthetic data allow models to achieve near-perfect performance on in-distribution tasks, they fail to support robust generalization to natural language. In contrast, models trained on natural sentences, systematically retrieved from Universal Dependencies, capture the

underlying patterns of passive alternation and generalize effectively across both natural and synthetic test sets. These findings highlight the enduring value of authentic, linguistically-grounded data for evaluating large language models: variability and complexity in natural corpora encourage models to learn abstract structural generalizations rather than memorizing surface patterns. Structured and annotated natural data remain crucial for rigorous linguistic evaluation and for probing the true syntactic and semantic knowledge of pretrained models cross-linguistically.

Limitations

Future work could address the limitations of this contribution by expanding language coverage and alternation phenomena, exploring additional models and architectures, and performing comprehensive validation, as well as a human upperbound.

Ethics

We used datasets derived from publicly available corpora, which may include content such as news articles and other publicly accessible materials. It is important to note that these datasets may contain sensitive or potentially upsetting topics. We acknowledge that such content could be distressing to some individuals. We encourage users to approach the results with awareness of these considerations.

Acknowledgements

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

8. Bibliographical References

Lorenzo Alfieri and Fabio Tamburini. 2016. (Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-IT 2016)*, pages 19–23, Napoli, Italy.

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.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siconatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. [UniMorph 4.0: Universal Morphology](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The Evalita 2014 Dependency Parsing Task. In *Proceedings of CLiC-it 2014 and EVALITA 2014*, pages 1–8, Pisa, Italy. Pisa University Press.

Joan Bresnan. 2016. [Lifetime Achievement Award: Linguistics: The Garden and the Bush](#). *Computational Linguistics*, 42(4):599–617.

Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2022. [Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian](#). *Frontiers in Psychology*, Volume 13 - 2022.

- Luigi Burzio. 1986. *Italian Syntax: A Government and Binding Approach*. Reidel, Dordrecht.
- Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abidin. 2024. [On the Diversity of Synthetic Data and its Impact on Training Large Language Models](#). *arXiv preprint*, arXiv: 2410.15226.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.
- Marco Civico. 2025. [Linguistic statistical universals: comparing computer- and human-generated texts](#). *International Journal of Digital Humanities*, 7(1):1–37.
- Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.
- Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.
- Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. [Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.
- Edward Gibson and Evelina Fedorenko. 2013. [The Need for Quantitative Methods in Syntax and Semantics Research](#). *Language and Cognitive Processes*, 28(1–2):88–124.
- Anita Goyal and Samira Mahmoud. 2024. [A Systematic Review of Synthetic Data Generation Techniques for NLP](#). *Electronics*, 13(17):3509.
- Thomas L Griffiths, Jian-Qiao Zhu, Erin Grant, and R Thomas McCoy. 2024. Bayes in the Age of Intelligent Machines. *Current Directions in Psychological Science*, 33(5):283–291.
- Bruno Guillaume. 2021. [Graph Matching and Graph Rewriting: GREW tools for Corpus Exploration, Maintenance and Conversion](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.
- Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en universal dependencies. *Traitement Automatique des Langues*, 60(2):71–95. HAL: hal-02267418.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint*, arXiv:2501.12948.
- Paul J Hopper and Sandra A Thompson. 1980. Transitivity in Grammar and Discourse. *Language*, pages 251–299.
- Chunyang Jiang, Giuseppe Samo, Vivi Nastase, and Paola Merlo. 2024. [BLM-It - Blackbird Language Matrices for Italian: A CALAMITA Challenge](#). In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1135–1143, Pisa, Italy. CEUR Workshop Proceedings.
- Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. [Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore. Association for Computational Linguistics.
- Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. *arXiv preprint arXiv:2504.02768*.
- 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.
- Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. [TSNLP - Test Suites for Natural Language Processing](#). In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.
- Cara Su-Yi Leong and Tal Linzen. 2023. [Language models can learn exceptions to syntactic rules](#). In *Proceedings of the Society for Computation in Linguistics 2023*, pages 133–144, Amherst, MA. Association for Computational Linguistics.

- Beth Levin. 1993. *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.
- Paola Merlo. 2023a. [Blackbird language matrices \(BLM\), a new task for rule-like generalization in neural networks: Can large language models pass the test?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8119–8152, Singapore. Association for Computational Linguistics.
- Paola Merlo. 2023b. [Blackbird language matrices \(BLM\), a new task for rule-like generalization in neural networks: Motivations and Formal Specifications](#). *arXiv*, arXiv: 2306.11444.
- Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. 2025. [Synthetic data generation using large language models: Advances in text and code](#). *IEEE Access*, 13:134615–134633.
- Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024a. [Exploring Italian sentence embeddings properties through multi-tasking](#). In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 620–630, Pisa, Italy. CEUR Workshop Proceedings.
- Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024b. [Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement](#). In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 631–643, Pisa, Italy. CEUR Workshop Proceedings.
- Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2025. [Multilingual vs. Monolingual Transformer Models in Encoding Linguistic Structure and Lexical Abstraction](#). In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 826–836, Cagliari, Italy. CEUR Workshop Proceedings.
- Malvina Nissim, Danilo Croce, Viviana Patti, Pierpaolo Basile, Giuseppe Attanasio, Elio Musacchio, Matteo Rinaldi, Federico Borazio, Maria Francis, Jacopo Gili, Daniel Scalena, Begoña Altuna, Ekhi Azurmendi, Valerio Basile, Luisa Bentivogli, Arianna Bisazza, Marianna Bolognesi, Dominique Brunato, Tommaso Caselli, Silvia Casola, Maria Cassese, Mauro Cettolo, Claudia Collacciani, Leonardo De Cosmo, Maria Pia Di Buono, Andrea Esuli, Julen Etxaniz, Chiara Ferrando, Alessia Fidelangeli, Simona Frenda, Achille Fusco, Marco Gaido, Andrea Galassi, Federico Galli, Luca Giordano, Mattia Goffetti, Itziar Gonzalez-Dios, Lorenzo Gregori, Giulia Grundler, Sandro Iannaccone, Chunyang Jiang, Moreno La Quatra, Francesca Lagioia, Soda Marem Lo, Marco Madeddu, Bernardo Magnini, Raffaele Manna, Fabio Mercorio, Paola Merlo, Arianna Muti, Vivi Nastase, Matteo Negri, Dario Onorati, Elena Palmieri, Sara Papi, Lucia Passaro, Giulia Pensa, Andrea Piergentili, Daniele Poterti, Giovanni Puccetti, Federico Ranaldi, Leonardo Ranaldi, Andrea Amelio Ravelli, Martina Rosola, Elena Sofia Ruzzetti, Giuseppe Samo, Andrea Santilli, Piera Santin, Gabriele Sarti, Giovanni Sartor, Beatrice Savoldi, Antonio Serino, Andrea Seveso, Lucia Siciliani, Paolo Torroni, Rossella Varvara, Andrea Zaninello, Asya Zanollo, Fabio Massimo Zanzotto, Kamyar Zeinalipour, and Andrea Zugarini. 2025. [Challenging the Abilities of Large Language Models in Italian: a Community Initiative](#). *arXiv preprint*, arXiv: 2512.04759.
- Joakim Nivre. 2015. [Towards a Universal Grammar for Natural Language Processing](#). In *International conference on intelligent text processing and computational linguistics*, pages 3–16. Springer.
- Juri Opitz, Shira Wein, and Nathan Schneider. 2025. [Natural language processing relies on linguistics](#). *Computational Linguistics*, 51(3):1009–1032.
- Mattia Proietti, Gianluca Leboni, and Alessandro Lenci. 2022. [Does BERT Recognize an Agent? Modeling Dowty’s Proto-Roles with Contextual Embeddings](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4101–4112, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Malka Rappaport Hovav and Beth Levin. 2024. [Variable Agentivity: Polysemy or underspecification](#). *Glossa: a journal of general linguistics*, 9(1).
- John C. Raven. 1938. [Standardization of progressive matrices](#). *British Journal of Medical Psychology*, 19:137–150.
- Tanya Reinhart. 1981. [Pragmatics and linguistics: An analysis of sentence topics](#). *Philosophica*, 27.
- Giuseppe Samo and Paola Merlo. 2021. [Intervention effects in clefts: a study in quantitative computational syntax](#). *Glossa: a Journal of General Linguistics*, 6(1):145.
- Giuseppe Samo and Paola Merlo. 2026a. [Datasets for Verb Alternations across Languages: BLM Templates and Data Augmentation Strategies](#). *arXiv*, arXiv: 2603.15295.
- Giuseppe Samo and Paola Merlo. 2026b. [Modelling the morphology of verbal paradigms: A](#)

- case study in the tokenization of Turkish and Hebrew. In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 82–94, Rabat, Morocco. Association for Computational Linguistics.
- Giuseppe Samo, Vivi Nastase, Chunyang Jiang, and Paola Merlo. 2023. [BLM-s/IE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12276–12287, Singapore. Association for Computational Linguistics.
- Ryohei Sasano, Daisuke Kawahara, Sadao Kurohashi, and Manabu Okumura. 2013. [Automatic knowledge acquisition for case alternation between the passive and active voices in Japanese](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1213–1223, Seattle, Washington, USA. Association for Computational Linguistics.
- Job Schepens, Hanna Woloszyn, Nicole Marx, and Benjamin Gagl. 2025. [Can Large Language Models Generate Useful Linguistic Corpora?: A Case Study of the Word Frequency Effect in Young German Readers](#). *Open Mind*, 9:1597–1656.
- Carson T Schütze, Jon Sprouse, Robert J Podesva, and Devyani Sharma. 2013. Judgment data. *Research methods in linguistics*, pages 27–50.
- Djamé Seddah and Marie Candito. 2016. Hard time parsing questions: Building a questionbank for french. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.
- Anna Siewierska. 2013. [Passive constructions \(v2020.4\)](#). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.
- Jon Sprouse and Diogo Almeida. 2017. [Design sensitivity and statistical power in acceptability judgment experiments](#). *Glossa: a journal of general linguistics*, 2(1):14.
- Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. [BLiMP-NL: A Corpus of Dutch Minimal Pairs and Acceptability Judgments for Language Model Evaluation](#). *Computational Linguistics*, 51(4):1267–1301.
- Francesca Volpato. 2010. *The Acquisition of Relative Clauses and phi-Features: Evidence from Hearing and Hearing-impaired Populations*. Ph.D. thesis, Università Ca'Foscari Venezia.
- Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. [Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do Llamas Work in English? On the Latent Language of Multilingual Transformers](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Michael Wilson, Jackson Petty, and Robert Frank. 2023. [How abstract is linguistic generalization in large language models? experiments with argument structure](#). *Transactions of the Association for Computational Linguistics*, 11:1377–1395.
- Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. [CLiMP: A benchmark for Chinese language model evaluation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.
- Lingze Zhang and Ellie Pavlick. 2025. [Does Training on Synthetic Data Make Models Less Robust?](#) *arXiv preprint*, arXiv: 2502.07164.

Appendix

#ARGS	VOICE	ST	PATTERN	WITHOUT
2	Act	Q	V-[nsubj]-> Ag; V-[obj]-> Pat; Q [form="?"]	Y [upos=VERB]
1	Act	Decl	V-[nsubj]-> Ag; V-[obj]-> Pat	Q [form="?"]; Y [upos="VERB"]
2	Act	Q	V-[nsubj]-> Ag; Q [form="?"]	V-[obj]-> Pat; Y [upos="VERB"]
1	Act	Decl	V-[nsubj]-> Ag	V-[obj]-> Pat; Q [form="?"]; Y [upos="VERB"]
2	Pass	Q	V-[nsubj;pass]-> Pat; V-[obl:agent]-> Ag; Q [form="?"]	Y [upos="VERB"]
1	Pass	Decl	V-[nsubj;pass]-> Pat; V-[obl:agent]-> Ag	Q [form="?"]; Y [upos="VERB"]
1	Pass	Q	V-[nsubj;pass]-> Pat; Q [form="?"]	V-[obl:agent]-> Ag; Y [upos="VERB"]
1	Pass	Decl	V-[nsubj;pass]-> Pat	V-[obl:agent]-> Ag; Q [form="?"]; Y [upos="VERB"]

Table 1: Queries to retrieve natural instantiation example from UD treebanks with GREW-MATCH (Guillaume, 2021). #ARGS = number of arguments, ST = Sentence Type, Act = Active Voice, Pass = Passive Voice, Q = question, D = declarative.