‘Layer su Layer’: Identifying and Disambiguating the Italian NPN Construction in BERT’s family

Greta Gorzoni, Ludovica Pannitto, Francesca Masini
Alma Mater Studiorum - University of Bologna
greta.gorzoni@studio.unibo.it, ludovica.pannitto@unibo.it, francesca.masini@unibo.it

Abstract

Interpretability research has highlighted the importance of evaluating Pretrained Language Models (PLMs) and in particular contextual embeddings against explicit linguistic theories to determine what linguistic information they encode. This study focuses on the Italian NPN (noun$_i$-preposition-noun$_i$) constructional family, challenging some of the theoretical and methodological assumptions underlying previous experimental designs and extending this type of research to a lesser-investigated language. Contextual vector representations are extracted from BERT and used as input to layer-wise probing classifiers, systematically evaluating information encoded across the model’s internal layers. The results shed light on the extent to which constructional form and meaning are reflected in contextual embeddings, contributing empirical evidence to the dialogue between constructionist theory and neural language modelling.

Keywords: Construction Grammar, Contextual Embeddings, Interpretability in Language Models

1. Introduction

The remarkable empirical performance obtained by Pretrained Language Models (PLMs) across a wide range of tasks has fueled enthusiasm in both computational approaches and theoretical debates about language (Brown et al., 2020). Despite these successes, PLMs remain largely opaque (Rogers et al., 2020). High predictive accuracy does not automatically entail theoretical understanding: it remains unclear what kinds of linguistic information these models encode, how such information is internally structured, and to what extent their representations align with linguistically motivated categories.

Assessing a model’s linguistic competence therefore requires an explicit theoretical characterization of the phenomenon under investigation. Construction Grammar (CxG, Fillmore 1988; Goldberg 1995, 1996) offers a particularly suitable framework in this respect: within CxG, language is conceived as a structured network of conventionalized form–meaning pairings, namely Constructions (Cxns), which may vary in both complexity and schematicity. Since phenomena at the syntax–semantics interface remain comparatively underexplored in interpretability research (Graichen et al., 2026), a constructionist perspective provides a principled way to investigate whether contextual embeddings encode structured form–meaning associations rather than surface-level distributional regularities (Pannitto and Herbelot, 2023; Weissweiler et al., 2022; Rambelli, 2025).

This study focuses on the Italian NPN (noun$_i$-preposition-noun$_i$) constructional family (e.g., *layer su layer* ‘layer upon layer’, Masini 2024a) and builds on previous research on the English NPN Cxn (Scivetti and Schneider, 2025), which showed that BERT encodes information relevant to both Cxn identification and semantic disambiguation. By extending the investigation to Italian, where (as in English) NPN patterns form a network of horizontally related Cxns exhibiting semantic specialization and competition, this work tests whether constructional knowledge is testable across these different languages and across the NPN family of sister Cxns. Our work also addresses a theoretical need within CxG. While constructionist approaches provide a rich descriptive and explanatory framework, they increasingly call for empirical validation. Neural language models, whose representations emerge from large-scale distributional exposure, offer a unique opportunity to test usage-based constructionist hypotheses in a controlled and quantifiable way. By probing contextual embeddings for evidence of constructional organization, this work contributes to the dialogue between linguistic theory and neural language modelling, asking not just whether models perform well, but whether they encode linguistically meaningful structure.

The paper is structured as follows. Section 2 introduces the NPN construction, while Section 3 situates the study within the relevant literature. Section 4 formulates the research questions and details the methodological design. Section 5 introduces the dataset and experimental setup. Sections 6 and 7 present and discuss the results of the identification and semantic disambiguation experiments, respectively.[1]

[1] All code developed for this project is available on GitHub: https://github.com/GretaGorzoni00/NPN

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 203–220
16 May 2026.

2. The NPN Construction

NPN expressions challenge traditional grammatical categories and motivate a model capable of capturing phenomena along the lexicon–syntax continuum. Formally, the pattern consists of nominal reduplication interrupted by a preposition.

Construction Schema

$Noun_i$ Preposition $Noun_i$

Treating NPN expressions as semi-specified Cxns accounts for both their productivity and their idiosyncratic properties: nominal identity, determiner absence, restricted prepositional choice, as well as characteristic meanings such as succession, distributivity, or contact, which are encoded within the constructional schema itself rather than derived from syntactic operations.

Despite his complex positioning in the constructionist framework (Goldberg, 1996), in his influential analysis of the English NPN pattern, Jackendoff (2008) identifies a limited set of productive prepositions (*by, for, to, after, upon*) and associates them with a restricted inventory of core meanings (e.g., *succession*, *juxtaposition*, *comparison*), alongside numerous idiomatic instantiations. Within this account, NPN is treated as a unified abstract constructional schema pairing a formal template with a family of related meanings. This view has been challenged by Sommerer and Baumann (2021), who argue that the NPN family does not instantiate a single highly abstract Cxn. Instead, they propose a network of horizontally linked, semi-schematic Cxns, emphasizing usage patterns and semantic specialization over maximal abstraction.

The Italian NPN family has been investigated by Masini (2024a), who adopts a similar horizontal-network perspective. Based on corpus evidence from CORIS[2], Masini identifies 8 semi-schematic Cxns, summarized in Table 1. Her dataset amounts to 1,298 types with varying token frequencies (Masini, 2024b). This study focuses only on Cxns with the prepositions *a* 'at/to' and *su* 'on'.

The Italian NPN family exhibits substantial overlap in the noun lexemes licensed by different patterns (Masini, 2024a). Such overlap and semantic competition make the Italian NPN family a particularly informative test case for contextual embeddings, as the coexistence of partially overlapping Cxns with subtly differentiated semantic profiles raises the question of how such distinctions are encoded in distributional representations and whether they are accessible to PLMs.

[2] *Corpus di Riferimento per l'Italiano Scritto*; Favretti et al. 2002.

	Form	Meaning	Types	Tokens
1	$[[x]_{N_i}\ a\ [x]_{N_i}]_{\text{MOD}_j}$	Succession, iteration, distributivity	21	399
2	$[[x]_{N_i}\ a\ [x]_{N_i}]_{\text{MOD}_j}$	Juxtaposition, contact	27	1108
3	$[[x]_{N_i^{sg}}\ su\ [x]_{N_i^{sg}}]_{\text{MOD}_j}$	Succession, iteration, distributivity	71	178
4	$[[x]_{N_i^{pl}}\ su\ [x]_{N_i^{pl}}]_{\text{MOD}_j}$	Greater plurality, accumulation	252	403
5	$[[x]_{N_i}\ per\ [x]_{N_i}]_{\text{MOD}_j}$	Succession, iteration, distributivity	385	3324
6	$[[x]_{N_i}\ per\ [x]_{N_i}]_{\text{SUB.CL}_j}$	Inescapable presupposition	21	21
7	$[[x]_{N_i}\ dopo\ [x]_{N_i}]_{\text{MOD}_j}$	Succession, iteration, distributivity	368	2387
8	$[[x]_{N_i}\ contro\ [x]_{N_i}]_{\text{MOD}_j}$	Juxtaposition, contact	56	170

Table 1: The 8 Italian NPN Cxns postulated by (Masini, 2024a). This work deals with Cxns 1–4.

3. Related work

Recent work has investigated whether LLMs encode constructional knowledge using a variety of experimental designs. One line of research (Tayyar Madabushi et al., 2020; Tayyar Madabushi and Bonial, 2025) examines multiple Cxns organized along a gradient of schematicity, testing whether models generalize across instantiations and whether increasingly schematic patterns remain accessible in contextual embeddings. A complementary approach focuses on individual Cxns and operationalizes constructional knowledge through controlled discrimination tasks, often contrasting target Cxns with carefully designed distractors (Scivetti and Schneider, 2025; Weissweiler et al., 2022). Recently, attention has turned to rare but productive Cxns (Weissweiler et al., 2025), which provide a stringent test for constructional generalization: because they are infrequent, successful modelling cannot rely solely on collostructional frequency effects but must capture construction-specific contributions.

Results converge on a graded pattern: lower-level, lexically anchored Cxns tend to be more accessible to LLMs, while schematic patterns pose greater challenges (Bonial and Tayyar Madabushi, 2024). At the same time, recent work has questioned core assumptions underlying constructional interpretability. Jumelet et al. (2024) argue that model generalizations should not only be measured in terms of performance but also compared to human generalization patterns. Dunn and Eida (2025) further caution against confirmation bias in Cxn probing, emphasizing the importance of detecting false positives and avoiding uncritical expansion of the Constructicon (viz., the network of Cxns) based on model behaviour alone.

This work directly builds upon Scivetti and Schneider (2025), who investigate the English NPN Constructional pattern using a probing framework. In their study, contextual embeddings extracted from BERT are used as input to a linear classifier to assess whether constructional information is encoded in the model's representations. Their dataset includes instances of the NPN Cxn (e.g., 'I was living *moment to moment*') contrasted with superficially similar distractors (e.g., 'In Rome largesse was doled out by *individuals to individuals*'), with lemma-level disjoint train–test splits to prevent lexical memorization. Two experiments target Cxn identification and one addresses semantic disambiguation. Results show that BERT-based probes reliably distinguish Cxns from distractors and capture semantic distinctions, with peak performance in middle-to-late layers, outperforming static embedding baselines.

4. Research Questions and Methodological Design

As Cxns are assumed to be inherently language-specific, probing constructional knowledge requires moving beyond the English-centric focus that characterises much of the existing literature. Moreover, the NPN Cxn occupies an intermediate position on the lexicon–syntax continuum, making it a suitable test case for assessing whether models can capture constructional generalizations that are neither fully schematic nor fully lexically specified.

This study addresses two research questions:

RQ1 Can BERT distinguish instances of the Italian NPN Cxn from distractors?

RQ2 Can BERT distinguish between the construction-specific meanings associated with different NPN instantiations within the Italian NPN family?

Following Scivetti and Schneider (2025), we adopt a probing framework in which contextual embeddings extracted from BERT family's models are used as input to a linear classifier. The preposition is operationally treated as the structural head of the Cxn (Jackendoff, 2008), and its embedding is selected as a primary locus for probing.

To address the research questions, two complementary tasks are implemented: (i) Cxn identification, testing whether the model distinguishes actual NPN instances from formally similar sequences (see Section 6), and (ii) semantic disambiguation, assessing whether the information encoded in the contextual embedings are adequate to distinguish between the construction-specific meanings associated with different NPN instantiations.

Our study extends Scivetti and Schneider (2025) in several important ways. First, the dataset includes different horizontally related Italian NPN semi-schematic Cxns featuring prepositions *a* 'at/to' and *su* 'on' (namely Cxns 1, 2, 3 and 4 in Table 1): this design tests whether embeddings capture constructional generalizations within a network of sister Cxns, with localized competition (Masini, 2024a).

Second, the distractors set is broadened, including distinct Cxns (e.g., PNPN) identified in the literature (see Example 1: similar cases are excluded from Scivetti and Schneider (2025)'s setup) beyond surface-isomorphic patterns: this refines the identification task moving from syntactic discrimination to identification of true form–meaning pairings.

(1) [...] *con una successione da* ***agenzia ad agenzia*** *quasi automatica.*
'[...] with an almost automatic succession from **agency to agency**'.

Third, both the embedding of the `[UNK]` token (used as prepositional substitute) and the embedding of the preposition itself (henceforth, `PREP`) are probed, allowing us to compare abstracted and lexically grounded representations, and to assess how constructional information interacts with prepositional semantics.

5. Methods

5.1. Data

The dataset used in this study (Gorzoni et al., 2026) is derived from the Italian NPN dataset presented in Masini (2024a), extended with full sentential contexts extracted from CORIS[3].

The full dataset contains 3,256 attested instances of the Italian NPN constructional pattern instantiated by the prepositions *a* 'at/to' and *su* 'on'. Following the annotation schema proposed in (Masini, 2024a), each occurrence is manually annotated with one of five semantic labels: *succession/iteration/distributivity*, *greater_plurality/accumulation*, *juxtaposition/contact*, *connection/transition*, and *idiosyncratic*. The dataset further comprises 1,751 distractor instances spanning eight pattern types, for a detailed description see Appendix A. All data were manually cleaned to remove ill-formed sentences. Since the present study focuses exclusively on Cxns (1), (2), (3) and (4) in Table 2, only Cxn instances annotated with the *succession/iteration/distributivity*,

[3]The manually annotated dataset of Italian NPN Constructions and Distractors (Gorzoni et al., 2026), including semantic annotations and inter-annotator agreement data, is archived on Zenodo: https://zenodo.org/records/18268135

greater_plurality/accumulation and *juxtaposition/-contact* label were included in the operative dataset for the analysis. The resulting dataset consists of 1,281 constructional instances and 989 distractors.[4] All instances in the dataset were annotated for their *meaning*, the *lemma* instantiating the Cxn and its *number*. A subset of 100 Cxns (equally balanced across Cxns (1), (2), (3) and (4) in Table 1) were cross-annotated by a group of 5 annotators. Annotation quality of Cxns' meaning was assessed using Cohen's κ and Krippendorff's α (Artstein and Poesio, 2008). Pairwise Cohen's κ shows strong to near-perfect reliability across annotator pairs ($\kappa = 0.79 - 0.91$). Nominal α is high ($\alpha = 0.858$) and further increases when a reduced penalty is assigned to confusions between semantically adjacent labels (*succession/iteration/distributivity* and *greater_plurality/accumulation*, $\alpha = 0.892$). The entire dataset was filtered to only retain longer sentences (> 5 tokens) and at most 30 items with the same lemma per preposition.

For each experiment, data is then split into 5 training and test partitions using an 80/20 ratio. Unlike Scivetti and Schneider (2025), we do not enforce full lemma-level disjointness across splits. Instead, we adopt a modified strategy: lemma–label pairs are never shared between training and test sets, while allowing the same lemma to occur across labels. This design preserves lexical separation at the level of individual labels while retaining cross-label lexical overlap. In particular, it enables evaluation on cases where a lemma appears in training only with the opposite label, providing a stricter test of whether constructional distinctions are recoverable from contextual embeddings beyond lexical identity. Class balancing is applied to both training and test data. Tables B.1–B.4 summarize the composition of the data splits used for both experiments (i) and (ii).

5.2. Models and probing

Probing methods aim to investigate which types of information are encoded in a model's internal representations by training an auxiliary classifier, commonly referred to as a probing classifier. The representations are first extracted from a pre-trained model and then used as input for the probing classifier, together with labels corresponding to a linguistic property that has been explicitly operationalised. The central assumption underlying probing is that the probing classifier's performance reflects the extent to which the target property is accessible in the representations. Crucially, to draw meaningful conclusions about the presence of linguistic information in the embeddings, the probing setup must rely on a deliberately weak classifier paired with a sufficiently complex task (Hewitt and Liang, 2019). This constraint helps ensure that high performance is attributable to information encoded in the representations themselves, rather than to the probe's capacity to learn the task independently.

To evaluate whether constructional information is linearly accessible in contextual embeddings, we train a separate logistic regression probing classifier on embeddings extracted from each layer of the four considered BERT family's models, and track their performance across layers. Scivetti and Schneider (2025), we adopt a BERT-base architecture[5] (Staatsbibliothek and Schweter 2025, 12 layers, 768 hidden units, 12 attention heads) trained on Italian data. We also include multilingual BERT (mBERT[6], Devlin et al. 2019), that enables cross-linguistic comparison and allows partial replication of Scivetti and Schneider (2025) under a unified architecture. To further evaluate the role of monolingual specialization, we test UmBERTo[7] (Parisi et al., 2020), a RoBERTa-based model trained exclusively on Italian corpora: this allows to evaluate the effect of pretraining objectives. Finally, we include XLM-RoBERTa[8] (Conneau et al., 2019), a multilingual model trained on large-scale cross-lingual corpora, as a multilingual companion to UmBERTo. All models are used in inference mode without fine-tuning. Contextual embeddings are extracted layer-wise and evaluated using the same probing protocol across architectures. We consider representations corresponding to the prepositional token and its `[UNK]` substitute.

We also train a control classifier (Hewitt and Liang, 2019) with a random label assigned to each lemma, whose performance should be near chance given the absence of spurious correlations between train and test data. A probing linear classifier is also trained on static embeddings as baseline: differently from Scivetti and Schneider 2025, we adopt FastText (Bojanowski et al., 2017; Joulin et al., 2017) rather than GloVe as a static baseline because its subword-based representations are better suited to morphologically rich languages such as Italian, allowing us to control for lexical and inflectional variation.

5.3. Experimental setup

For the identification task, we perform binary classification (*Construction* vs. *Distractor*). For the disambiguation task, we perform multi-class classification across the three semantic labels. In order

[4] Numerous distractor instances belonging to the same distractor type in the full dataset were excluded in order to ensure heterogeneity.

[5] dbmdz/bert-base-italian-cased

[6] google-bert/bert-base-multilingual-cased

[7] Musixmatch/umberto-commoncrawl-cased-v1

[8] FacebookAI/xlm-roberta-base

to ensure maximum comparability with the previous study on English NPN constructions, we extended our experimental setup to the dataset introduced by (Scivetti and Schneider, 2025). Specifically, their data were converted into our format, subjected to our splitting procedure, and processed through the same embedding extraction and probing classification pipeline adopted for the Italian experiments. In addition, we computed a static baseline using English FastText vectors on the English configuration as well.

6. Identification task

The first experiment evaluates whether contextual embeddings extracted from BERT's models encode sufficient information to distinguish NPN constructions from distractors, and analyzes how the nature of the distractor patterns affect the probing classifier's behaviour.

In Scivetti and Schneider (2025)'s implementation, in fact, the identification task contrasts actual NPN instances with surface-isomorphic patterns. Because the distractor class is syntactically homogeneous and derives from a different structural configuration (e.g., the preposition does not function as the head in the distractor cases), the task may largely rely on syntactic structural cues, which could make it less clear to what extent specific constructional information is being actually tested.

Given the heterogeneous composition of our *distractor* portion of the dataset, three different sampling procedures were implemented[9]:

(i) a fully balanced configuration (henceforth, `SIMPLE`), where both training and test sets contain all distractor types;

(ii) a configuration where training is performed exclusively on structurally distinct constructions (i.e., PNPN, VERBAL and NSUNGIÙ) and testing on the full balanced test set (henceforth, `OTHER` configuration);

(iii) a configuration where, conversely, training is performed exclusively on surface-isomorphic patterns (i.e., *Thematic target*, *N-extended*, *NUM P NUM* and *Proper name inglobation* patterns), and evaluation on the full balanced test set (henceforth, `PSEUDO`).

The aim of these configurations is to investigate whether providing (`OTHER`), as opposed to withholding (`PSEUDO`), structurally informative negative examples that make constructional boundaries explicit has an impact on the performance of the probing classifier.

[9]All the the train test split and in particular the constraint needed for each configuration are menaged through the Python package (Pannitto, 2025)

With regard to configuration (i), as shown in Figure 1, probing classifiers trained on contextual embeddings, regardless of the specific model considered, achieve strong performance in distinguishing NPN Cxns from distractors, both for `[UNK]` and `PREP`.

As in Scivetti and Schneider (2025), the control classifier remains close to chance level, consistent with the selectivity criterion discussed by Hewitt and Liang (2019). The static lexical baseline performs well above chance: FastText probe provides a meaningful lower bound, which is outperformed by contextual embeddings only in middle to late layers, consistently with the interpretation that deeper layers encode more abstract information that is not available to static embeddings alone (Hewitt and Manning, 2019; Tenney et al., 2019; Scivetti and Schneider, 2025).

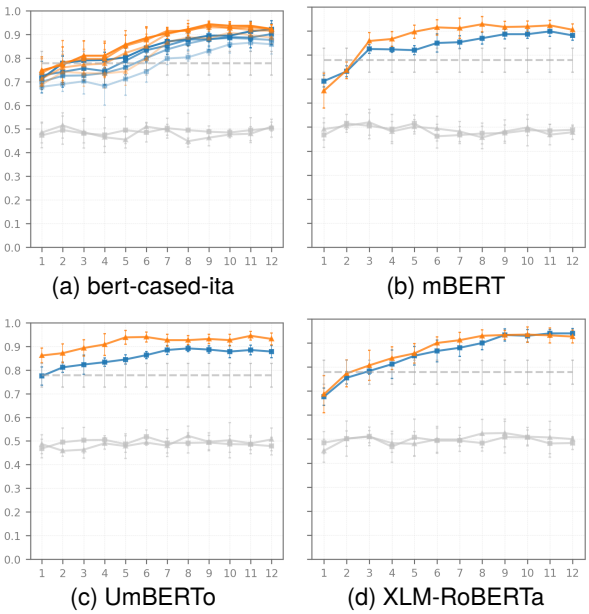


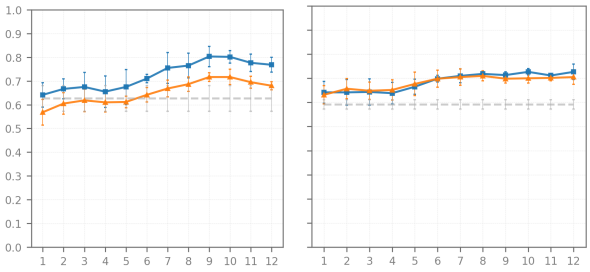
Figure 1: Accuracy of `[UNK]` (red lines, square dots) and `PREP` (orange lines, triangular dots) on Construction identification for the `SIMPLE` configuration. As in the following plots, the accuracy of the five probing classifiers resulting from the five random splits is averaged. Dashed grey line represents FastText baseline. Continuous grey lines refer to control classifiers. Figure (1a) includes decremental training configurations, line shading becomes progressively lighter as the number of training instances decreases (480 → 240 → 120 → 60). No substantial performance differences emerge across configurations. Figure C.1 in Appendix provides a qualitative visualisation of the embedding space across layers.

Results are confirmed on English NPN pattern as well: global performance is consistent both for `[UNK]` and `PREP` embeddings, but GloVe and FastText baselines perform quite differently on the task

(Figure 3).

Finally, the fact that [UNK] and PREP yield similar performance is informative. Replacing the preposition with [UNK] removes direct access to lexical information, yet the probe remains highly accurate. This suggests that successful discrimination does not crucially depend on access to the preposition's lexical semantics, but can be achieved on the basis of information contributed by the surrounding nouns and the broader sentential context. At the same time, the comparable performance of PREP suggests that lexical information about the preposition does not substantially increase accuracy beyond what is already available in the contextual configuration. These results support the conclusion that the probes capture information in BERT representations that reliably distinguishes the NPN Cxn from near-minimal distractor counterparts beyond what is possible through lexical semantic cues alone.

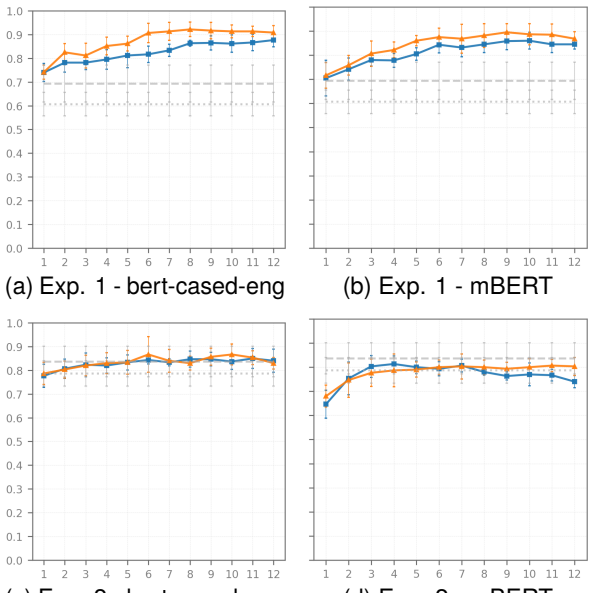
Configurations (ii) and (iii) yield slightly lower overall performance, while still consistently outperforming the baseline (Figure 2).



(a) OTHER (b) PSEUDO

Figure 2: Accuracy of [UNK] (red lines, square dots) and PREP (orange lines, triangular dots) on Construction identification for the OTHER and PSEUDO configurations. Dashed grey line represents FastText baseline.

The distribution of misclassifications (see Appendix D) highlights three main patterns, which are consistently observed across all models in the BERT family. In the SIMPLE configuration, all models exhibit very few classification errors, in line with the high overall accuracies. The vast majority of errors are False Positives, i.e. distractors incorrectly classified as constructions. This pattern suggests that constructional information is robustly encoded in the representations. The PSEUDO configuration illustrates how strongly the nature of the distractor instances defines the task itself. In this setting, the model is exposed to actual NPN constructions alongside distractors that are only superficially isomorphic but lack constructional status. In the test set structurally distinct constructions are more frequently misclassified. This suggests that, given the type of distractors encountered during training, the classifier might effectively learn to distinguish between constructional and non-constructional instances, rather than on deeper structural or construction-specific properties. In the OTHER configuration, PREP performance appears to be influenced by an imbalance in the training data, due to the constraints of this configuration. Since most distractors that instantiate a different Cxn are realised with the preposition *a*, instances with *su* are more frequently misclassified. This suggests that PREP representations are sensitive to prepositional frequency effects. By contrast, [UNK] representations do not seem to exhibit the same bias, indicating a more construction-oriented encoding less dependent on lexical skew. Taken together, these results highlight the importance of carefully defining what counts as a minimal pair for the Cxn under investigation, in order to faithfully translate the underlying theoretical question into a computational task. In probing experiments performance is strongly shaped by the experimental setup, in particular by the definition of distractors and minimal pairs. This suggests that probing results should be interpreted with care, as they reflect not only properties of the model, but also properties of the task design.



(a) Exp. 1 - bert-cased-eng (b) Exp. 1 - mBERT

(c) Exp. 2 - bert-cased-eng (d) Exp. 2 - mBERT

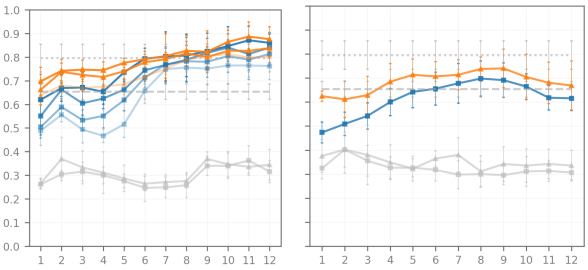
Figure 3: Accuracy of [UNK] (red lines, square dots) and PREP (orange lines, triangular dots) on both experiments on English data from Scivetti and Schneider (2025). Dashed grey line represents FastText baseline. Dotted grey line represents GloVe baseline.

7. Disambiguation task

Given the very high performance achieved in the experiment about the identification of NPN Cxn, ex-

tending the analysis beyond form, we now turn to examining the semantic dimension of the Cxn. Our setup is a multinomial three-class disambiguation problem: we only focus on the Cxn (1), (2), (3) and Cxn (4) in Table 1, which are associated to the three possible meanings of *juxtaposition/contact*, *succession/iteration/distributivity* or *greater_plurality/accumulation*.

Performance for both `[UNK]` and `PREP` embeddings, as shown in Figure 4, are above the baselines. In addition to the static baseline computed on the lemma of the reduplicated noun in the Cxn, this experiment includes a stronger baseline condition. Since one of the defining properties distinguishing Cxn (3) from (4) is the grammatical number of the reduplicated noun, we computed an additional FastText baseline using vectors built from the morphologically inflected noun: this proves substantially stronger than the lemma-based one. BERT models surpass it only in the highest layers, suggesting that lower layers do not encode information beyond what is already available from surface morphological cues, while higher layers capture more abstract constructional distinctions.



(a) bert-cased-ita (b) mBERT

(c) UmBERTo (d) XLM-RoBERTa

Figure 4: Accuracy of `[UNK]` (red lines, square dots) and `PREP` (orange lines, triangular dots) on Construction disambiguation task. Dashed grey line represents FastText baseline. Dotted grey line represents morphological FastText baseline. Continuous grey lines refer to control classifiers. Model (4a) includes decremental training configurations, line shading becomes progressively lighter as the number of training instances decreases (480 → 240 → 120 → 60). No substantial performance differences emerge across configurations. Figures C.2 and C.3, in Appendix, provide a qualitative visualisation of the embedding space across layers.

Error analysis for both `[UNK]` and `PREP` contextual embeddings reveal a highly structured pattern across the BERT family. The probing classifier achieves near-perfect performance in identifying the *greater_plurality/accumulation* and *succession/iteration/distributivity* meanings when instantiated with the preposition *su* 'on'. Although overall performance remains strong, some confusion emerges between *juxtaposition/contact* and *succession/iteration/distributivity* when both are instantiated with the preposition *a* 'at/to'. In other words, the model shows greater difficulty in distinguishing between semi-schematic Cxns that share the same surface preposition. Importantly, *greater_plurality/accumulation* is distinguished not only semantically but also through a salient surface morphological constraint (singular or plural number of the reduplicated noun), which may facilitate classification, as the the additional morphologically informed baseline suggests. Future work could isolate this formal property in a controlled design in order to assess the relative contribution of surface morphology and constructional semantics to probing performance.

A substantial proportion of the errors involves NPN Cxn (1) and (2), for which, respectively, labels *juxtaposition/contact* and *succession/iteration/distributivity* are predicted. These cases frequently involve high-frequency and highly conventionalised types, such as *gomito a gomito* ('elbow to elbow'), *petto a petto* ('chest to chest'), *fronte a fronte* ('forehead to forehead'), and spatial-configuration patterns such as *balcone a balcone* ('balcony to balcony'), *uscio a uscio* ('doorway to doorway'), and *porta a porta* ('door to door').

(2) *Qui lavorano i revisori **gomito a gomito** con i membri di Cosea*.
'Here auditors work **elbow to elbow** with members of Cosea[10]'
ANNOTATED: JUXT./CONTACT;
PREDICTED: SUCC./ITER./DISTRIBUTIVITY

(3) *Faccia a faccia e **petto a petto**, Big Jim serrò le mani sulle braccia di Andy e lo guardò negli occhi.*.
'Face to face and **chest to chest**, Big Jim closed his hands on Andy's arms and looked at him in his eyes.'
ANNOTATED: JUXT./CONTACT;
PREDICTED: SUCC./ITER./DISTRIBUTIVITY

Examples (2) and (3) show that, although the Cxn canonically encodes physical contact or spatial adjacency, the larger event structure evokes dynamic interaction unfolding in time (e.g. working sessions, fighting sequences). The probe may

[10]Cosea is an Italian company, active in the Emilia region.

therefore privilege an interpretation compatible with iterativity or sequential event structure over a purely configurational reading.

A similar oscillation emerges in spatial-distribution expressions.

(4) *Dopo anni passati prima come venditore* ***porta a porta*** *di assicurazioni e di calzature per donne (...) questo è il suo momento..*
'After years spent first as a **door to door** salesman of insurance policies and women's shoes (...) this is his moment.'
ANNOTATED: SUCC./ITER./DISTRIBUTIVITY;
PREDICTED: JUXT./CONTACT

In these cases, such as Example (4), the model appears to oscillate between an interpretation based on contact or adjacency between entities along a spatial path and an interpretation based on distribution over successive sites. The alternation suggests that, for entrenched lexicalised types, the boundary between *juxtaposition/contact* and *succession/iteration/distributivity* could be intrinsically gradient rather than categorical.

These results motivate a more focused examination of the *succession/iteration/distributivity* label, as it is compatible with multiple prepositions and does not rely on a strong surface constraint (unlike *greater_plurality/accumulation*). Its classification therefore provides a more stringent test of whether contextual embeddings capture abstract constructional meaning independently of lexical realisation.

7.1. Semantic generalisation

In order to create an experimental setup that allow us to test the probing classifier on unseen semi-schematic Cxns, we annotated 100 instances on Cxns (5) and (7) from Table 1 with sentential context extracted from CORIS: these have meaning of *succession/iteration/distributivity* and are realised by the prepositions *per* 'by' and *dopo* 'after'.

We trained the probing classifier on the full dataset comprising the three semantic labels together with the distractor instances, and then tested it exclusively on the newly annotated *per* 'by' and *dopo* 'after' instances. The composition of this dataset is shown in Appendix B.5.

Figure 5 reports accuracy across layers for `[UNK]` and `PREP` embeddings, together with static baselines. We can observe that `[UNK]` and `PREP` representations support robust generalisation to unseen prepositions, with performance reaching high accuracy in late layers. The task is intrinsically harder, as it is demonstrated by the drop in performance for both baselines. Nonetheless, results are consistent across different model in BERT family. `[UNK]` starts from moderate performance in the earliest layers and increases steadily, approaching 0.9 in the final layers. `PREP` shows a different

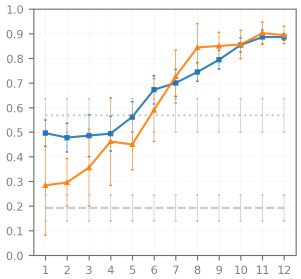


Figure 5: Accuracy of `[UNK]` (red lines, square dots) and `PREP` (orange lines, triangular dots) on Construction disambiguation. Dashed grey line represents FastText baseline. Dotted grey line represents morphological FastText baseline.

trajectory: it is markedly weaker and more variable in the lowest layers, where it performs close to baseline, but in 4–8 layers exhibits a sharp improvement, converging with `[UNK]` in late layers. This pattern suggests that early-layer `PREP` representations are strongly related to lexical information about the preposition, while higher layers increasingly integrate contextual information that can support a semantics-based decision even when the preposition itself is unseen in training.

These findings provide converging evidence that the semantic information exploited by the probing classifier is not merely a learned association between a small set of prepositions and labels. Instead, the late-layer gains for `[UNK]` and especially for `PREP` are consistent with an increasingly abstract encoding of constructional meaning, where the model can map unseen prepositional instantiations (*per* 'by', *dopo* 'after') onto a semantic category encountered in training (*succession/distributivity/iteration*). Under this interpretation, the experiment supports the conclusion that while lower BERT layers, particularly for `PREP`, remain more sensitive to surface lexical form, higher layers encode a representation rich enough to sustain generalisation across competing semi-schematic Cxns (Masini, 2024a) expressing *succession/distribution/iterativity* which can be realised by any of the prepositions *a* 'at/to', *su* 'on', *per* 'by', and *dopo* 'after'.

8. Conclusion

We presented two probing experiments addressing the identification and semantic disambiguation of Italian NPN constructions. To this end, we introduced an extended dataset including both constructional instances and carefully designed distractors, allowing for a controlled evaluation of construction-sensitive encoding.

We extended and enriched the setup of (Scivetti and Schneider, 2025) by comparing `[UNK]` and

PREP contextual embeddings, implementing alternative training configurations grounded in different minimal-pair and distractor definitions, and introducing a generalisation test on unseen prepositions instantiating unseen semi-schematic Cxns.

The aim of this study is to investigate whether the internal representations of models in the BERT family are compatible with a constructionist perspective, with respect to the Italian NPN constructional family. This question is addressed through a set of experiments targeting both form-related and meaning-related properties of the constructions.

The results suggest that distinguishing constructional patterns relies on the joint contribution of multiple factors, including lexical distribution and morphological cues. As indicated by the baseline performance, none of these factors in isolation is sufficient to fully resolve the task; rather, they appear to jointly support the information that is recoverable for classification.

The identification results show that constructional information is detectable in contextual embeddings. However, they also demonstrate that performance is highly sensitive to the composition of the training set: the nature of the distractors and the operationalisation of minimal pairs substantially shape the classification task. In this respect, careful interpretative caution is required, in line with recent observations on task sensitivity in probing studies (see also Dunn and Eida (2025)).

The disambiguation experiment further reveals that static embeddings constitute strong baselines. Lexical information, especially when coupled with morphological cues such as number of the reduplicated noun, proves highly informative and can account for a significant portion of the variance. Crucially, however, when the task involves generalisation to unseen semi-schematic constructions instantiating the same semantic value, contextual embeddings enable the classifier to achieve high accuracy precisely where static vectors show a marked performance drop. This suggests that higher layers encode more abstract constructional regularities that go beyond type-level lexical information.

Taken togheter, the results are consistent with a constructionist perspective, in that successful classification appears to rely on the joint contribution of multiple cues rather than on isolated lexical features.

Multilingual models (mBERT and XLM-R respectively) deserve specific attention: while their performance remain largely comparable to monolingual models in the identification task, the situation appears different for disambiguation. In disambiguation task, in fact, both mBERT and XLM-R underperform their monolingual counterpart. While the difference is negligible for English, the drop in performance is particularly evident for Italian data. This confirms the importance of extending interpretability investigations to languages other than English: taking a language-specific perspective like the one that CxG offers can help evaluating language-specific abilities of multilingual models.

Overall, the findings seem to provide empirically grounded support for an abstraction path in PLMs towards an higher-level construction linking the semi-schematic Cxns associated with the semantics of *succession/distribution/iterativity*, realized by the prepositions *a* 'at/to', *su* 'on', *per* 'by', and *dopo* 'after' within the NPN family.

At the same time, the results remain confined to the representational space within BERT's models and any theoretical implications must therefore be interpreted with caution.

In order to provide empirical grounding for the theoretical debate concerning the schematization of the NPN construction, the probing results are currently being cross-validated through behavioural experiments that directly compare human judgments with model predictions. This step will be essential to determine whether the computational evidence actually reflects cognitively plausible constructional abstractions and to more rigorously assess the theoretical validity of the proposed NPN schematization.

9. Limitations

The present study is subject to several limitations.

First, the analysis is restricted to a single constructional family, namely the Italian NPN Cxs. Although multiple prepositions (*a* 'at/to', *su* 'on', *per* 'by', *dopo* 'after') are included, they instantiate closely related constructions within the same constructional network, differing primarily in degree of productivity and conventionalisation. As such, the findings should be interpreted as evidence about this specific constructional domain rather than as a general account of constructional information in internal representations of the model. Extending the analysis to Cxns with different syntactic profiles and levels of schematicity would be necessary to assess the broader validity of the observed patterns. This is especially relevant given that NPN Cxns predominantly fulfil an adverbial function, which limits the diversity of syntactic configurations examined.

Second, the study focuses on models within the BERT family. While this allows for controlled comparison across variants, these models share core architectural and training properties, including encoder-only structure and a masked language modelling objective. Consequently, the results cannot be straightforwardly generalised to other classes of models. In particular, architectures with different training objectives or representational dynamics may encode constructional information dif-

ferently. Future work should therefore extend the analysis to decoder-only models, in order to assess whether the results observed here are specific to masked language modelling or reflect more general properties of transformer-based representations, including those shaped by next-token prediction.

Third, the methodology is limited to diagnostic probing. This design choice ensures comparability with prior work on the English NPN construction (Scivetti and Schneider, 2025), but it also constrains the scope of the conclusions. In particular, it remains unclear whether the information encoded in the model's internal representations and exploited by the probe is causally involved in the model's behaviour on the task. In other words, while probing suggests that the representational space can be analyzed in terms of constructional information, it does not establish that the model itself relies on this information when making predictions.

10. Ethics Statement

Annotators were recruited within an advanced Master's-level course as part of structured educational activities. Participation was entirely voluntary and had no impact on students' evaluation or academic standing. All participants were informed about the objectives of the study and the intended use of the collected data.

11. Bibliographical References

Ron Artstein and Massimo Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Claire Bonial and Harish Tayyar Madabushi. 2024. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Dunn and Mai Mohamed Eida. 2025. LLMs learn constructions that humans do not know. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 13–23, Düsseldorf, Germany. Association for Computational Linguistics.

R. Favretti, R. Rossini, F. Tamburini, and C. De Santis. 2002. CORIS/CODIS: A corpus of written italian based on a defined and a dynamic model.

Charles J. Fillmore. 1988. The mechanisms of "construction grammar". In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55, Berkeley. Berkeley Linguistics Society.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. The University of Chicago Press, Chicago.

Adele E. Goldberg. 1996. Jackendoff and construction-based grammar. *Cognitive Linguistics*, 7(1):3–20.

Greta Gorzoni, Ludovica Pannitto, and Francesca Masini. 2026. NPN Construction and Distractor dataset.

Nora Graichen, Iria de Dios-Flores, and Gemma Boleda. 2026. The grammar of transformers: A systematic review of interpretability research on syntactic knowledge in language models.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ray Jackendoff. 2008. "Construction after Construction" and Its Theoretical Challenges. *Language*, 84(1):8–28.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Jaap Jumelet, Lisa Bylinina, Willem Zuidema, and Jakub Szymanik. 2024. Black big boxes: Do language models hide a theory of adjective order?

Francesca Masini. 2024a. Costruzioni su costruzioni: idiomaticity and regularity of npn discontinuous reduplications in italian. In *I fraseologismi schematici. Questioni descrittive e teoriche*, pages 51–82. Aracne, Roma.

Francesca Masini. 2024b. Npn discontinuous reduplications in italian: Dataset.

Ludovica Pannitto. 2025. sample-dataset.

Ludovica Pannitto and Aurélie Herbelot. 2023. CALaMo: a constructionist assessment of language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 21–30, Washington, D.C. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

Giulia Rambelli. 2025. *Constructions and Compositionality: Cognitive and Computational Explorations*. Cambridge University Press.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Wesley Scivetti and Nathan Schneider. 2025. Construction identification and disambiguation using bert: A case study of npn. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376. Association for Computational Linguistics.

Lotte Sommerer and Andreas Baumann. 2021. Of absent mothers, strong sisters and peculiar daughters: The constructional network of english npn constructions. *Cognitive Linguistics*, 32(1):97–131.

Bayerische Staatsbibliothek and Stefan Schweter. 2025. bert-base-italian-cased (revision 843e404).

Harish Tayyar Madabushi and Claire Bonial. 2025. Construction grammar evidence for how LLMs use context-directed extrapolation to solve tasks. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 190–201, Düsseldorf, Germany. Association for Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882. Association for Computational Linguistics.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2025. Hybrid human-llm corpus construction and llm evaluation for the caused-motion construction. *Northern European Journal of Language Technology*, 11(1):27–57.

Appendix

A. Distractors

PNPN - Construction - (*a*)

Description:
The PNPN Construction $[P\ N_i\ P\ N_i]$ encodes a meaning of *iterativity*, *succession*, and *transition* between elements belonging to the same semantic category. In this Construction, the nominal slots are occupied by the same noun, repeated in a discontinuous form, while the preceding preposition —typically *a* or *da* — specifies the direction or type of relation between the two occurrences.

Example (ID 720): *Il tutto tra l'altro avviene nel giro di pochi mesi con una successione **da agenzia ad agenzia** quasi automatica*
'All of this, moreover, happens within a few months with an almost automatic succession **from agency to agency**'.

Verbal - Construction - (*a*)

Description:
The verbal Construction $[V\ N_i\ a\ N_i]$ is realised by a restricted set of transitive verbs that allow both a direct object and a prepositional complement introduced by *a*. Among the most frequent are verbs of accumulation or transfer, such as *aggiungere* ('to add'), *unire* ('to join'), *sommare* ('to sum'), and *accostare* ('to bring together'). This Construction exhibits a high degree of semantic conventionalisation: the identity of the nouns generates an intensifying or cumulative meaning that goes beyond the literal meaning of the verb.

Example (ID 1027): *Quel fatto **aggiunse irritazione a irritazione**.*
'That fact **added irritation to irritation**'.

NUM P NUM - isomorphic pattern — (*a*)

Description:
This pattern expresses a relationship of **equivalence** or **balance** between two numerical values that are identical or opposed (identical, in the case of superficial identity with NPN)

Example (ID 1498): *Il risultato non si sbloccava e si era sempre sullo **zero a zero**.*
'The result did not get unlocked and it was always at **zero to zero**'.

N extended - isomorphic pattern - (*a*, *su*)

Description:
These instances superficially replicate the NPN Construction, but one or both occurrences of the noun are part of a larger complex phrase or multiword expression and do not constitute an autonomous instance of the same syntactic category. Consequently, they do not convey the iterative, distributive, or juxtapositional relations that define actual NPN constructions.

Example (ID 229): *Le due telecamere alternavano **primi piani a piani a figura intera** da angolazioni diverse.*
'The two cameras alternated **close-ups to full shots** from different angles'.
Example (ID 2193): *La Commissione può ora rimborsare **spese generali di base su base forfetaria** o versare importi forfetari per piccoli progetti.*
'The Commission can now reimburse basic overhead costs on a flat-rate basis or pay lump-sum amounts for small projects'.

Proper name inglobation - isomorphic pattern — (*su*)

Description:
Accidental lexical repetition in which one or both noun occurrences belong to a proper name. In such cases, the repetition has no constructive or syntactic function, resulting merely from the contiguous presence of the same lemma in two elements of the sentence.

Example (ID 2189): *La posizione del **Comune di Arezzo su Arezzo Fiere** è chiarissima.*
'The position of the Municipality of Arezzo on Arezzo Fiere is very clear'.

N su N giù - Construction - (*su*)

Description:
Combines nominal repetition with the directional pair *su–giù* ('up, down'), producing a rhythmic and iterative effect. It typically occurs in colloquial or performative contexts, marking repetition or cyclic succession of events or actions, but can also convey an iterative or ironic meaning, evoking repetition and excess in a rhythmically salient way. The su–giù pair contributes to a sense of saturation and cyclical movement.

Example (ID 2174): *Come se ci fosse bisogno di una voce dall' alto che dica **"bandiere su bandiere giù"**.*
'As if there were a need for a voice from above saying 'flags up, flags down' '.

Num P Num - isomorphic pattern — (*su*)

Description:
Expresses a proportional or evaluative relation between two numerical values, typically used in quantificational or performative contexts such as voting, scores, or results. The preposition *su* introduces the reference term, establishing a part-whole or ratio relation ("x obtained out of y possible"). When the two numerals are identical, the expression acquires idiomatic and highly conventionaliased value and is especially productive in sports and journalistic language.

Example (ID 2127): *Per nostra fortuna sono stati fecondati **3 su 3**, il 100%, pronti per essere impiantati in utero.*
'Fortunately for us, 3 out of 3, 100'%, were fertilized, ready to be implanted in the uterus'.

Thematic target - isomorphic pattern — (*su*)

Description:
The surface structure N-su-N is preserved, but the function does not correspond to the canonical meanings of the Construction (plurality, iteration, accumulation, distributivity, or connection). These distractors feature a second noun specifying the thematic domain.

Example (ID 2088): *Comunque non spetta a questa introduzione dire delle nostre tre autrici: qui si tratta piuttosto di chiedersi perché, a voler parlare di romanzo popolare in Italia tra i due secoli, si sono scelte tre donne che scrivono per **donne su donne**.*
'Anyway, it is not the role of this introduction to say anything about our three authors: here, it is rather a question of asking why, when one wants to speak of popular novels in Italy between the two centuries, three women who write for women on women were chosen'.

B. Training and Test sets

Label	Prep	Train	Test
CXN	*a*	120	30
CXN	*su*	120	30
CXN (total)	–	240	60
DISTR	*a*	120	30
DISTR	*su*	120	30
DISTR (total)	–	240	60
total	–	**480**	**120**

Table B.1: SIMPLE training test split configuration

Label	Prep	Train	Test
CXN	*a*	70	30
CXN	*su*	70	30
CXN (total)	–	140	60
DISTR	*a*	25	30
DISTR	*su*	115	30
DISTR (total)	–	140	60
total	–	**280**	**120**

Table B.2: PSEUDO training test split configuration

Label	Prep	Train	Test
CXN	*a*	55	30
CXN	*su*	55	30
CXN (total)	–	110	60
DISTR	*a*	105	30
DISTR	*su*	5	30
DISTR (total)	–	110	60
total	–	**220**	**120**

Table B.3: OTHER training test split configuration

Label	Prep	Train	Test
Succession	*a*	60	15
Succession	*su*	60	15
Succession (total)	–	120	30
Accumulation	su	120	30
Juxtaposition	a	120	30
total	–	**360**	**90**

Table B.4: Semantic disambiguation train–test split configuration.

Label	Prep	Train	Test
Succession	*a*	30	–
Succession	*su*	30	–
Succession	*per*	–	50
Succession	*dopo*	–	50
Succession (total)	–	60	100
Accumulation	*su*	60	–
Juxtaposition	*a*	60	–
Distractor	*a*	30	–
Distractor	*su*	30	–
Training set (total)	–	**360**	**100**

Table B.5: *per* and *dopo* train–test split configurations.

C. PCA

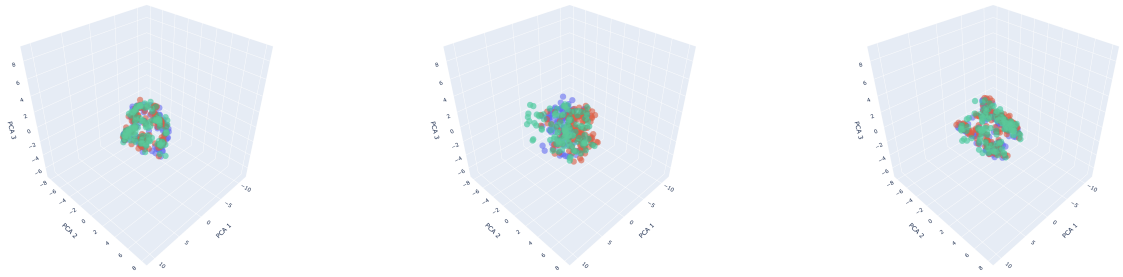
(a) Layer 1 (b) Layer 2 (c) Layer 3

(d) Layer 4 (e) Layer 5 (f) Layer 6

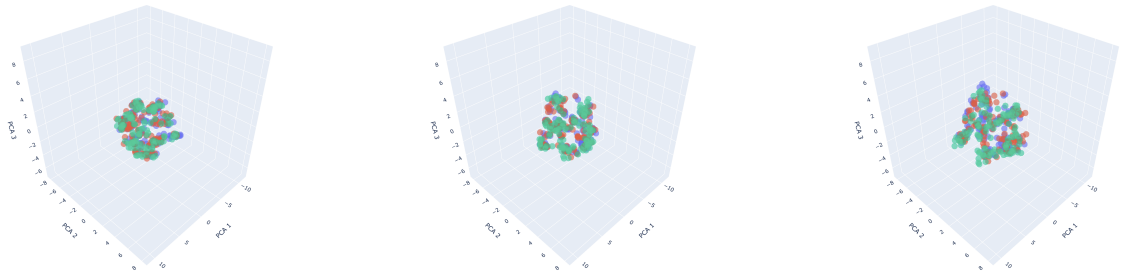
(g) Layer 7 (h) Layer 8 (i) Layer 9

(j) Layer 10 (k) Layer 11 (l) Layer 12

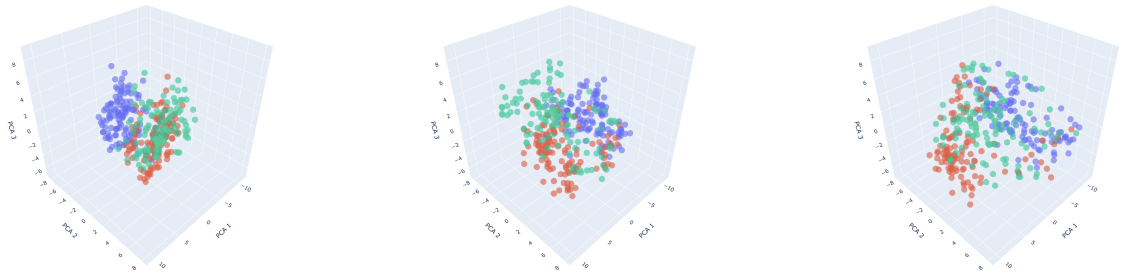
Figure C.1: PCA-based three-dimensional projection of UNK BERT embeddings across layers for NPN constructions (red) and distractors (red). An animated version of the visualisation is available at https://gretagorzoni00.github.io/NPN_contextual_embeddings/.



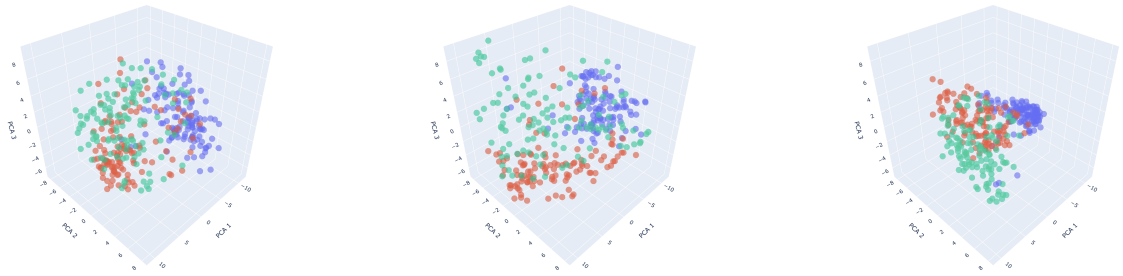
(a) Layer 1 (b) Layer 2 (c) Layer 3



(d) Layer 4 (e) Layer 5 (f) Layer 6

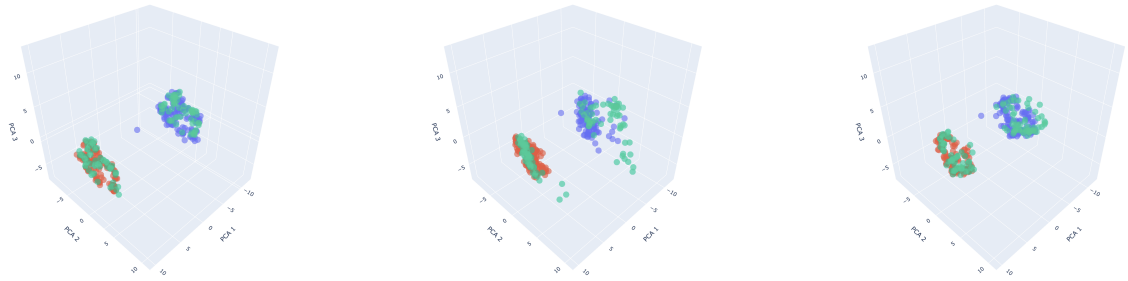


(g) Layer 7 (h) Layer 8 (i) Layer 9

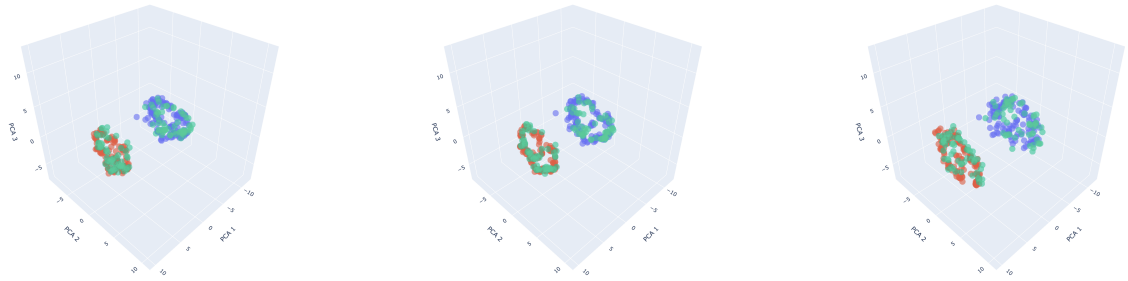


(j) Layer 10 (k) Layer 11 (l) Layer 12

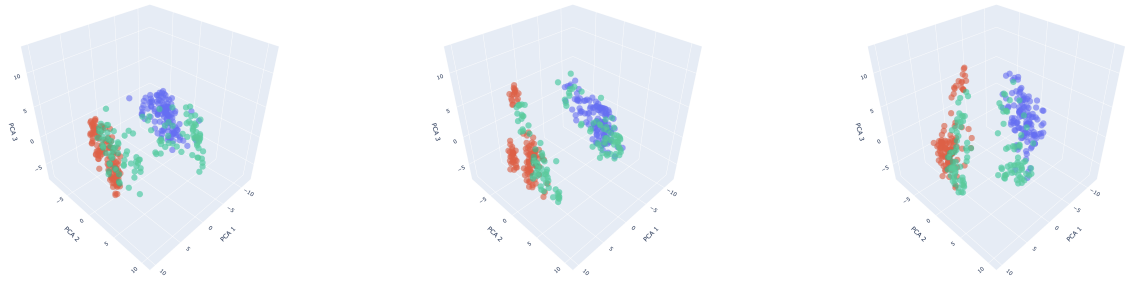
Figure C.2: Three-dimensional PCA projection of UNK-based BERT contextual embeddings across layers for NPN semi-schematic constructions. Data points are colour-coded by semantic value: *juxtaposition/-contact* (red), *greater_plurality/accumulation* (red), and *succession/iteration/distributivity* (green). The visualisation is based on a balanced subset of 360 instances. An animated version of the visualisation is available at https://gretagorzoni00.github.io/NPN_contextual_embeddings/.



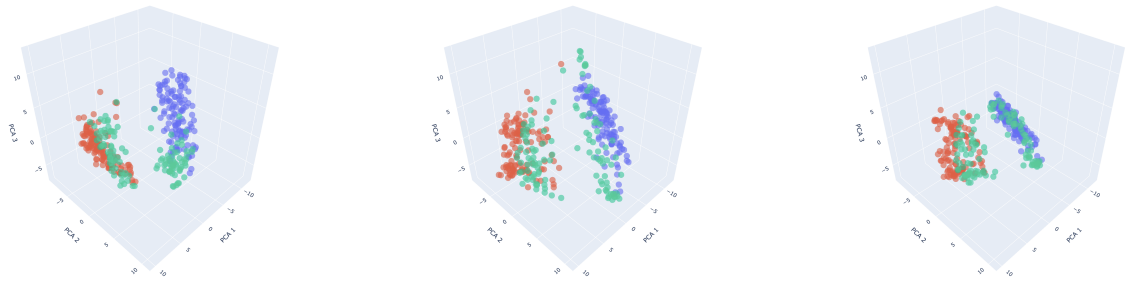
(a) Layer 1 (b) Layer 2 (c) Layer 3



(d) Layer 4 (e) Layer 5 (f) Layer 6



(g) Layer 7 (h) Layer 8 (i) Layer 9



(j) Layer 10 (k) Layer 11 (l) Layer 12

Figure C.3: Three-dimensional PCA projection of PREP-based BERT contextual embeddings across layers for NPN semi-schematic Constructions. Data points are color-coded by semantic value: *juxtaposition/contact* (red), *greater_plurality/accumulation* (red), and *succession/iteration/distributivity* (green). The visualisation is based on a balanced subset of 360 instances. An animated version of the visualisation is available at https://gretagorzoni00.github.io/NPN_contextual_embeddings/.

D. Confusion matrices

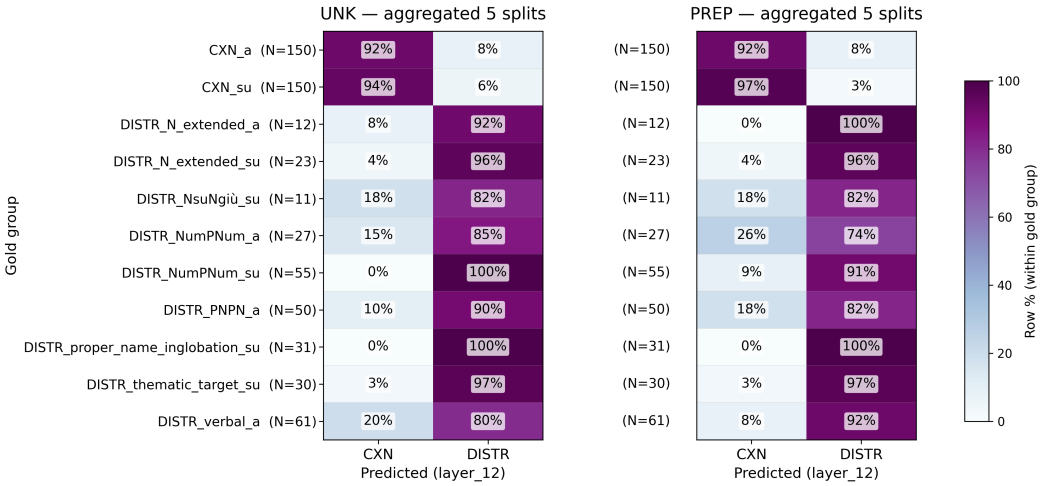


Figure D.1: Confusion matrices for the `SIMPLE` configuration in the Cxn identification experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.

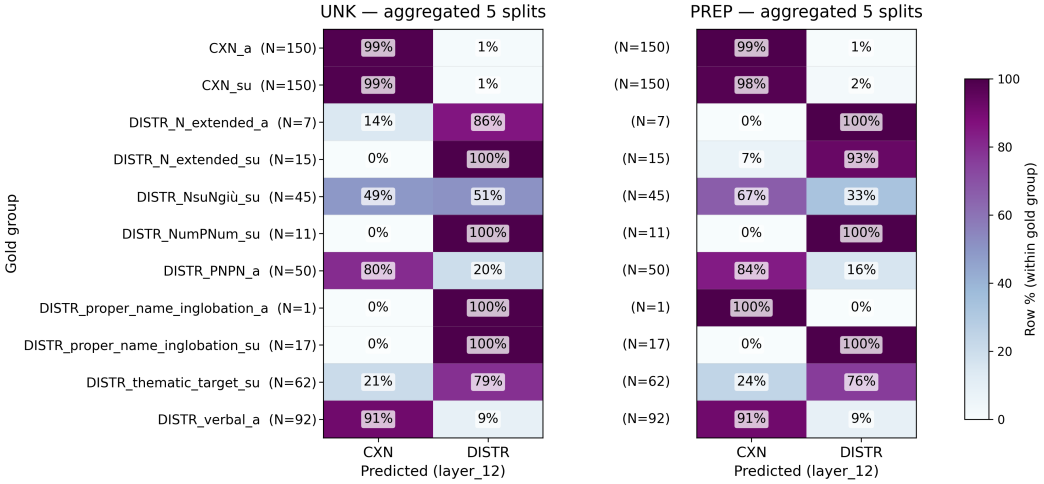


Figure D.2: Confusion matrices for the `PSEUDO` configuration in the Cxn identification experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.

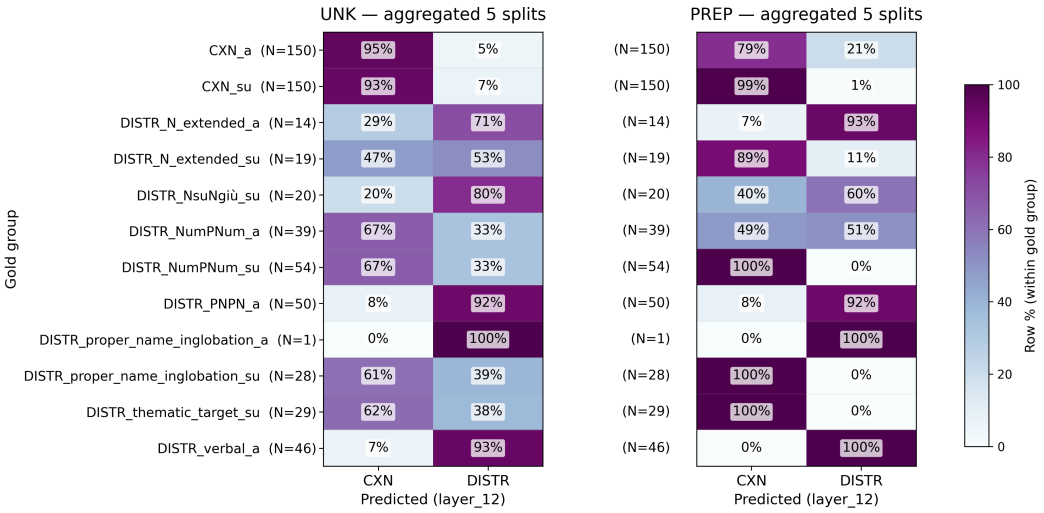


Figure D.3: Confusion matrices for the `OTHER` configuration in the Cxn identification experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.

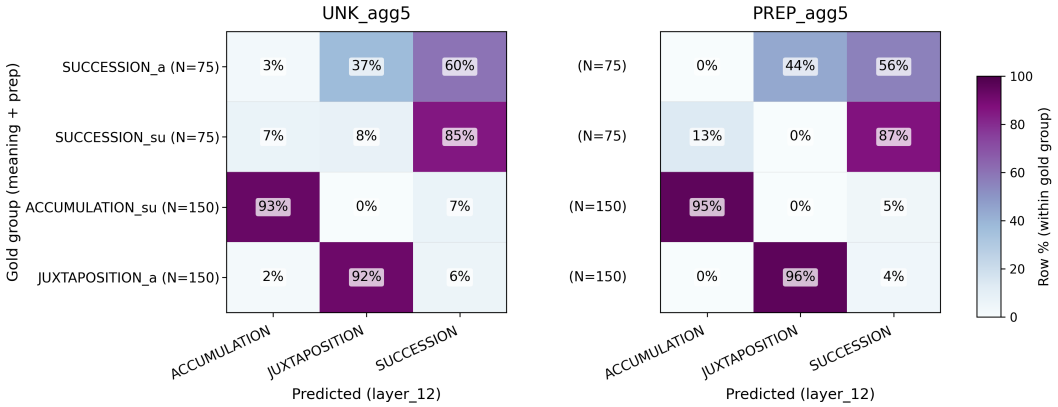


Figure D.4: Confusion matrices for the Cxn disambiguation experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.